**Reinforcement Learning approaches for penetration testing**

Table of Contents

[Declaration i](#_Toc95842141)

[Dedication ii](#_Toc95842142)

[Acknowledgements iii](#_Toc95842143)

[Table of Contents 1](#_Toc95842144)

[List of Figures 3](#_Toc95842145)

[List of Tables 4](#_Toc95842146)

[List of Abbreviations 5](#_Toc95842147)

[List of Symbols 6](#_Toc95842148)

[Title of Paper 7](#_Toc95842149)

[I. Introduction 7](#_Toc95842150)

[Background 7](#_Toc95842151)

[Aim of Research 7](#_Toc95842152)

[Structure of Paper 7](#_Toc95842153)

[II. Approach 7](#_Toc95842154)

[III. Experiments / Testing / Measurements / Data 7](#_Toc95842155)

[IV. Analysis and Discussion of Results 7](#_Toc95842156)

[V. Conclusions 7](#_Toc95842157)

[VI. Future Work 7](#_Toc95842158)

[VII. Instructions 7](#_Toc95842159)

[Formatting 7](#_Toc95842160)

[Title 7](#_Toc95842161)

[Headings 7](#_Toc95842162)

[Detailed instructions 7](#_Toc95842163)

[Indicative Length 7](#_Toc95842164)

[Use of colour 7](#_Toc95842165)

[Figures and Tables 8](#_Toc95842166)

[References 8](#_Toc95842167)

[Remove additional information 8](#_Toc95842168)

[Appendix A: Literature Review 9](#_Toc95842169)

[Introduction of Literature Review 9](#_Toc95842170)

[Body of Literature Review 9](#_Toc95842171)

[Conclusions from the Literature Review 9](#_Toc95842172)

[References of the Literature Review 9](#_Toc95842173)

[Indicative Length of Appendix A: Literature Review 10](#_Toc95842174)

[Appendix B: Data and Programs Used for this Report 11](#_Toc95842175)

[Required 11](#_Toc95842176)

[Optional 11](#_Toc95842177)

[Appendix C: Figures 12](#_Toc95842178)

[Appendix D: Tables 13](#_Toc95842179)

[References 14](#_Toc95842180)

List of Figures

*(required, include page numbers)[[1]](#footnote-1)*

Figure 1, Title of Figure #

List of Tables

*(required if you use tables, include page numbers)[[2]](#footnote-2)*

Table 1, Title of Table #

List of Abbreviations

*(required if you use abbreviations)[[3]](#footnote-3)*

List of Symbols

*(required if you use symbols)[[4]](#footnote-4)*

Selection of Reward in Reinforcement Learning for Auto penetration Testing

*Abstract*—Penetration testing has a main role to identify the security of network security and becomes more and more important as the interconnections by networks are increasing rapidly. Penetration testing needs network experts who have full knowledge and rich experiences in network attack and security, however it becomes more and more difficult to find such experts. Recently, as AI has rapid growth in all parts of science and technology, auto penetration testing methods based on AI have been developing. Especially, many researchers have interesting in reinforcement learning based auto penetration testing. There are many papers and packages for auto penetration testing with Reinforcement learning. However most of them concentrate on implementation of reinforcement learning model for penetration testing. To get optimal performance by reinforcement learning in penetration testing, it is very important to set the reward function correctly. In this paper, we develop the deep reinforcement learning based penetration testing method using efficient reward function.

*Keywords*—Cyber Security, Network Attack, Penetration test, Reinforcement learning, Reward,

# Introduction

The introduction to your paper should go here. This section usually has the following subheadings.

## Background

[1] In recent, the computer networks have been deployed with latest security checks & controls to overcome the external computer threats. However numbers of Computer threats have been increased with more complexity as well as Sophistication and frequency of those threats. It is important to deploy Cyber Security for safeguarding the computer systems and individual and governments’ important data [1]. One of the major practices being done nowadays is to do penetration testing for evaluating the current security of the system and recommend for major upgrades or controls required. Penetration testing is an established mechanism to check the security of digital systems, varying from an individual application to a network of digital communications. It doesn’t include only websites or software application, but all hardware and network setups too. In Penetration testing, an authorized attack is performed to find the security vulnerabilities that could be used by a Hacker to overcome to exploit the system.

Penetrating testing is approved as an effective method to find the security vulnerabilities of the system as it is a simulation of what would be performed by an attacker. However, the effectiveness of this method comes with one drawback that it requires high skills and time to perform such penetration testing.

After Introduction of European General Data Protection Regulation, the Penetration Testing has become a mandatory and crucial component for evaluating the Cyber security threats of the systems. As the size of digital systems and complexity is increasing, it is impacting the cost and requirements of performing such tests and therefore it is increasing demand of Such Security Testing & Quality Assurance professionals by the big Security Testing firms across the globe. [2] The Penetration testing is a multiple stages process which requires highly professional and skilled resources because of complexity of medium and large networks {2}.

It is important to mention here that penetration testing and security functional testing are two different terms and perform their functions in different ways. The Security Functional Testing can describe the Systems security controls’ correct behavior while Penetration Testing is used to calculate the level of difficulty one needs to penetrate an organization’s security controls without authorizations and access its information.

In Penetration Testing, an unauthorized user will attack the system with either Manual or Automated Tools for assessing the security controls and in possible circumstances where they can be disabled. Through Penetration Testing, the Security Professionals can address the security layers and possible vulnerability points and highlight the vulnerabilities of low, medium and high-severity in a system. The Businesses and their operations gets high benefits from valued quality testing tools like Penetration testing. For a Business Organization, the Penetrating testing is mandatory because it prevents the theft of financial data, complying the industry security requirements, reliable information to both stakeholders and the customers. [4].

The Penetrating Testing is an effective method to evaluate the strength of existing security products and major upgrades or actions required in such resources. The Systems’ penetration testing provides the “existence of an issue” and a valid case for the organization management to make investment in overcoming it. [5].

The One of leading Information Technology(IT) and networking hardware and services Firm CISCO SYSTEMS estimated in 2015 that there were 1 Million System Security Jobs unfilled in the world. Currently, a number of such security checking tools have been developed that help to perform penetration testing and improve the security of the systems. Metasploit framework is such popular tool that was in development from 2003 is now available for Penetration testing.[8]. Metasploit includes a library of known security threats to perform for your system penetration testing. Also, it includes many useful tools to scan the information gathering on a target place. The tools like Metasploit helps a Penetration Tester to perform testing at high level rather analyzing and finding the vulnerabilities at low levels with his manually defined checklists. The Penetration testing tools help the Testers to work quickly with high level of reliability, while it creates a friendly and easy-to-use assessment method for entry level testers. ​[9]. While such Testing Tools have been beneficial to the Cyber security testing industry, the problem lies in the availability of professional trainers to perform such operations. Also, the methods of manually testing the security of systems become a burden with increasing complexity of the system’s growth. A new trending approach to perform system testing with complexity of system structures and growth is to apply such tests from the Artificial Intelligence (AI) in order to automate the testing process and get quick outcomes, which are not possible with manually testing with the tools. The reliability and efficiency of testing has one approach where Artificial Intelligence the techniques for the process automation and it has took an original concept of “Concept Graph” where an already designed computer network is utilized to launch a Test Attack and it shows the current threats and vulnerabilities in the existing system. ​[10]​. To learn the possible ways in a ways in which a system hacker can enter the system through a security breach can be learned through Attack Graphs. Such Attack Graphs need sound knowledge of the system to use them and find the exploits that is different or unrealistic from attacker point of view or such attack being performed in the real world. Further, it needs to construct a manual attack graph for each system being tested.

The Partially Observable Markov Decision Process is an another approach that is used for the similar systems’ testing. In Partially Observable Markov Decision Process, the attack on a system is modeled in an assumption attacker exploits are partially added in the testing process and the perform his exploits and assuming the simulation process that such, while it models and writes the threats in its configuration during this simulation process. ​[11], [12]. Due to computing mechanism of POMDPs, this approach can be performed on an individual machine but it doesn’t scale perfect in a big environment [12]​. The more research work is required to design methods that should automate the pen testing for handling the threats and exploits possible in a system with their computational properties. A possible solution to overcome such problems is to use MDB approach. In this approach, the configuration of host computer status is ignored and possibility of each possible attack is tested. [13]​.

The MDP method is computationally more feasible as compared to POMDPs approach because it doesn’t require the understanding and knowledge of hosting and network configurations. The MDP approach also requires a previous knowledge of the system because it understands each target computing machine as same instead of defining possible targets for each machine and it requires success possibilities for every performed action. It affects the accuracy of selecting the right actions for every host because this method doesn’t utilize the knowledge of host and network configuration. Another similar approach to address the problems without knowing the host and network properties is Reinforcement Learning [14]. The state representations are the set of actions which determine the actions achievements by a Reinforcement Learning Agent and these are performed and reward function for the RL Agent. By communicating with its environment, the RL Agent learns about the actions’ policy. The Reinforcement Learning has gained a high popularity while it was used in to beat the agents of World GO Champion. As compared to other testing procedures, the Reinforcement Learning has not been commonly adopted but in robotic tasks the RL has been successfully adopted ​[15]​.

In cases where detailed model of the environment is not known, a general approach is used by Reinforcement Learning model. It can define the tasks in a reward function by selecting the actions which should be performed while Automated Penetrating Testing. The problem lies with ever changing state of exploits and threats in network environment and with on-going changes in the software and updates, It becomes difficult to manage an accurate and updated model for such executing any of the actions.

There are certain challenges while adopting the RL Model because it required large volume of data to understand and learn a policy of performing best actions, method for Automated System Testing. The RL Agent is trained by the data gathered through the simulations.

Recently, it is important to know that RL based penetration methods gets more and more attracts.

The Numbers of packages for auto penetration testing using deep reinforcement learning are developed. The Mulval is a tool used to analyse the network security vulnerabilities and it improves the visualization by deeply trimming on logic graph.[]

NAS is a python package which is given in the format of OpenAI gym Environment.[]

It is additionally planned to model network pen-testing at a better level of abstraction in order to permit fast prototyping and testing of algorithms. The network model defines the organization, connection and configuration of machines on the network and is defined by the tuple subnetworks, topology, hosts, services, firewalls. This model seeks to abstract away those aspects of a real world network that aren’t needed when creating autonomous agents, such as specific types of machine connections and the location of switches and routers. The objective for this abstraction is to keep the simulator as simple as possible while still working at the level that the agent is expected to work at, allowing it to decide which scans or exploits to use against a certain system and in what order. This basic network model is also employed to keep it as generic and scalable as possible.

[] and [] developed deep Reinforcement learning algorithm based on NAS.

They get the result for small networks.

However the number of states and actions will be increased rapidly when the numbers of hosts are increasing.

In Table 1, the states and number of actions are given

This produces serious problems to apply RL method to the penetration testing since training can’t be converged if the numbers of states are too big.

To make the training converge with high numbers of states, it will need high computation and time.

Preceding papers mainly concentrate on how to apply RL methods to penetration testing.

To apply RL algorithm to penetration testing on big networks, we should find the method to reduce the time that training to be converged.

## Aim of Research

In this paper, we are going to check the preceding Deep RL based penetration testing methods and find the solution to apply Deep RL method to big networks.

Firstly, we analysed the environment and RL algorithm in the preceding methods and find the problems.

Secondly, we should propose new method to solve the problems.

To do it, updating for Reward will be concentrated.

How to set Rewards for actions are important element to decide the performance of training and test of RL however preceding papers don’t concentrate on this.

In this paper, we are going to propose new deep RL based auto penetration testing method and compare the performance s.

## Structure of Paper

This paper constitutes 5 chapters.

Section 1 shows introduction to the paper.

Section 2 shows the approach for this paper.

Subsection 2.1 describes the NAS environment, test and analyse DQN agent on NAS.

Subsection 2.2 describes the updating DQN agent of NAS based on the results of subsection 2.1.

Section 3 represents the experiments for proposed algorithm and shows the results.

Section 4 analyses the results of experiments and show the performance of new algorithm.

Section 5 contains the conclusions for this paper.

Appendix contains tables, figures and programs used

in this paper.

# Approach

As described in introduction, there exists some auto pen-testing methods using deep Reinforcement learning and also several packages.

Here, we develop DQN based auto pen-testing methods using NAS which is a python package to generate environment for pen-testing. NAS generates gym Environment so we can easily use deep learning python packages such as tensorflow, pytorch and openAI which are widely used to.

This section has two sub sections.

First part is to test the deep Q Reinforcement Learning based Auto penetration testing method on tiny and big networks and analyse the performance of them.

We use the DQN agent which is given in [] and get the sources of it on [https://github.com/Jjschwartz/NetworkAttackSimulator](https://github.com/Jjschwartz)​.

Second part is to update the penetration testing based on the analysis for first part.

1. **Analysis of DQN agent based penetration testing**

In this sub-section we tested DQN based penetration testing methods on two scenarios and analysed the performance.

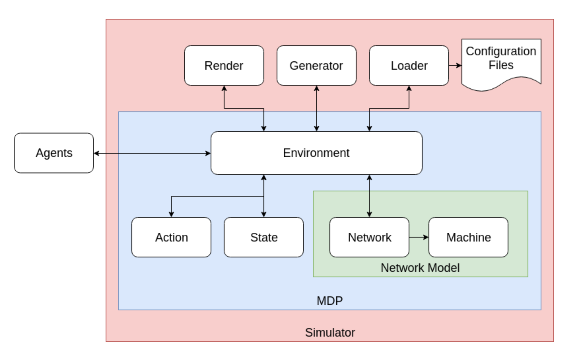


Figure 2-1. Structure of penetration testing based on RL in NAS.

Figure 2-1 shows the structure of penetration testing based on RL in NAS. As we can see from the figure, penetration testing is constituted of two main parts, ie. agents and environment.

Agents receives the network information from environment and attack the network based on it.

Environment is given by NAS and consistitute several parts. Network model is used to simulate network environment and has two parts, ie. n network such as information of subnetworks, links between subnetworks and so on. Machine simulates the host with os, service and process. The action consists of a scan action and a deterministic or non-deterministic exploit for each service and machine on the network. State is the set that contains the state of each machine on the network.

Loader is to load scenario file which has network information for subnetworks and hosts.

Generator creates openAI gym environment from scenario file or benchmark scenario name.

Render is a library to show the state of environment using figures and shell. Detailed description for NAS environments are given in the following subsections.

* 1. **Environment**

1. ***Network Model***

In Network Model, the Organization’s network including server machine, workstations, firewalls, networks & Subnetworks, the configuration and connectivity of the machines defined by a tuple. In Figure 2-2, an Example of Network Model including installed firewall security, 16 machines and five subnetworks with firewall is shown. In this network model, the number of service to run has been defined at machine level so it can operate any number of the services. The further details about it have been mentioned in below paragraphs.

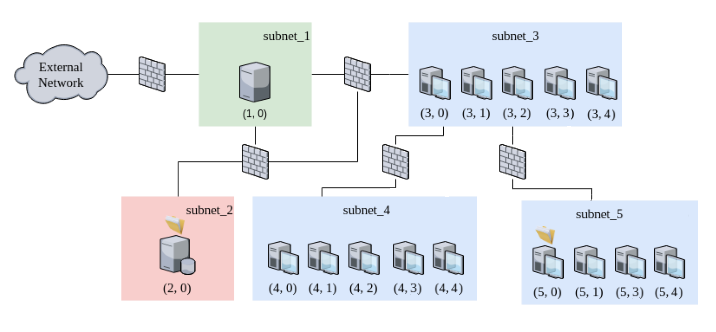


Figure 2-2. Example of network model.

This model presents a real time network model where details are not required in terms of locations of routers, switches and the machines. The major reason to use such simulation model is because it is simple and easy to learn presentation of how agent will work in a simulation and which exploits or scans should be performed at specific machines connected in this network. Also, this model determines in which order the scans and exploits on the machines and the network. While adopting this practice in high fidelity networks, the details of which ports to use for communication are handled by implementation of application in that environment. If the exploit works in a specified scenario, it is launched while knowing about all the details on the lower level of this exploit testing​[18]. In this network model for Exploit testing, the defined details for performing the actions are used so that it shall be easily scaled to the network map and to keep it as a standard method.

* **Machines**

In a Network Model, the machines are the most important building block. In NAS, the Machines refers to any type of device which is available on a network and which can be exploited through communicating the attacks in the network. Every machine while available on network has a unique Machine ID and Subnet ID along with its configuration. The Machines are defined with different value while performing an attack as machine with high value is the one that owner wants to protect or Attacker wants to exploit to the low value machine which is not important in terms information available in it.

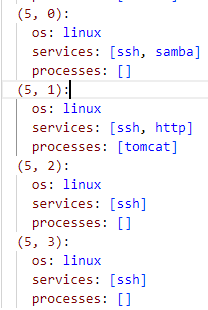


Figure 2-3. Examples of machine definitions.

In **Figure 2-3**, the examples of Machines definition are given. The services running on every machine in a network can be communicated with other machines on the same network or other network but with the firewall or subnet giving permission to access it. It is not necessary that each machine on a network should have same configuration and it depends on services available for all of the machines. The different machines available on a network have difference purposes to serve and therefore their configuration varies, e.g. Server Machines, Workstations, File storage servers and web servers. Since the hackers are interested in exploiting the services running on those machines, the vulnerability of each machine differs and level of security needed too.

* **Services**

After Machines, the second major component included in a network model are the services running on the machines and their level of security as well as configuration is different as defined above. The Services are including any type of Software Application that are running on the Machines and communicating in the network. They may be compared as software that are communicating with software and devices available on the port. The Services are vulnerability points in a machine network and attacker may know about specific exploits that can hack the machine by accessing its services communication in NAS.

In real world scenarios, the networking protocols only follow to keep track of services which are vulnerable or important to attacks and ignore the services which may not vulnerable. An agent’s job is to find the services running on each machine and apply the exploits it is vulnerable to as in above reasoning for the NAS, we assume that one action can exploit the each service. In Network Map, each services has been defined with a unique ID, the cost of applying the threat and a possibility percentage that tester will be successful in apply the threat.

**Figure 2-3** shows how an attacker has certain exploits for the HTTP, SSH, FTP services in a network, while Figure **2-4** shows cost of exploits, the success probability of each exploit on these services and a set of services which can be exploited. It is not necessary that each service should have an actual environment name but rather a unique ID, as shown in the figure. It is so that NAS shall generate IDs for the services with any numbers of services and machines in a defined test environment. However, the services names are used while performing the same configuration in real world scenario instead of defining IDs for the services. The services’ names (e.g. Samba Version 4.1.0) are used in real environment for the patching purposes and to specify the vulnerability to that version of the application.

* **Sub-Networks**

The Sub-Network are an important component in a network map. There are multiple subnets or sub networks in every network.

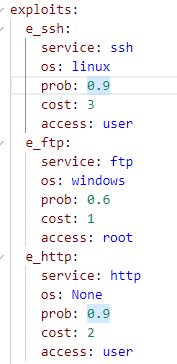


Figure 2-4. Examples of exploitable services on a network.

The machines in a network are connected on a smaller network that is called subnet and it may include a group of machines for communicating with each other in a larger network. The machines on a subnet have their own Subnet Address which is mentioned in its network address (e.g. 3 in (3,0)) is a subnet. It is simple method to define the IP Addresses which have 32-bit subnet mask or 32-bit string for defining the machine address, subnet or the network. In NAS Network Model, we use simple System as it is single network working as compared to Use of IP address which is used on Public networks for communicating to millions of the machines. The communication in a subnet is managed by firewall configured and the network topology, meanwhile machines on a subnet can fully communicate with each other while they may be restricted on other subnets on the same network.

* **Topology**

The Topology in a network map shows how subnets on are controlled and connected to other subnets and can directly communicate among themselves are with the subnets on the other networks. **In Figure 2-2,** a network topology is shown where the Subnets 1, 2 and 3 are connected to each other and can free communicate data on this subnet while Subnet 1 has been configured to communicate with external subnets and networks. So, an attacker may get access to the Subnet 1 to communicate and apply threats /exploits on the other Subnets (2 and 3) and a whole setup of machines in this network or his target point machine at this network. The subnets may be viewed as vertices with connection as edges while understanding the concept of this network topology.

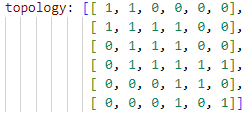


Figure 2-5. Example of topology definitions

Figure 2-5 show the example of topology definitions for the scenarios given in Figure 2-2. As we can see in the example, topology is given by the adjacent matrix which has the values of 0 for not connected links and 1 for connected links.

* **Firewall**

An important and key security component of a Network is Firewall. Firewalls are the components which exist in connections between the subnets as well connectivity of machines in network or external networks in an environment. The firewalls define which services to be accessed or controlled while machines are installed in a network. Firewalls control the number of services accessible on a certain subnet or a network. The firewalls allow selected services to be available on a subnet or accessible to another subnet from one machine to another, while blocking a certain subnet. The access of traffic to control in a subnet or associated subnets as blocked or permitted is defined in Firewall set of rules.

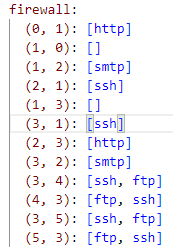


Figure 2-6 . Example of firewall configuration in the network

In Figure 2-6, configuration of firewalls for the network given in figure 2-2 is given. The first number of each pair is the source subnetwork and second is destination subnetwork.

The accessibility is defined on ports however the rules of firewall on subnet have been defined on the services rather than the network ports. The Firewalls are important because only selected ports can communicate the data or access a certain system.

* Scenarios

Environments described above are implemented using scenario files. Each Scenario file has the format of yaml and can configure all parts of networks. The scenario files which are used for testing are given in Appendix.

Following figures show the structures of networks used for testing.

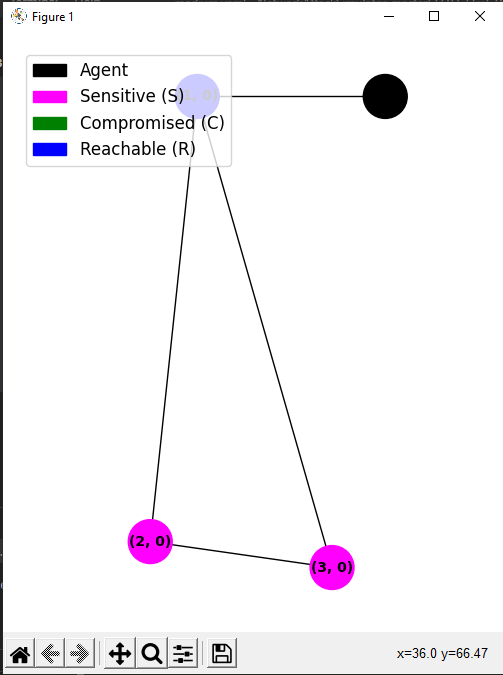


Figure 2-7. Tiny Scenario

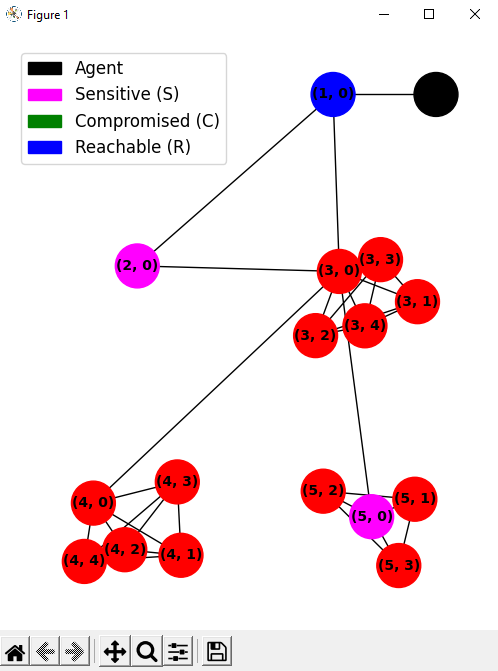


Figure 2-8. Medium Scenario

Tiny scenario contains three subnetworks and three hosts.

Medium scenario contains 5 subnetworks and 16 hosts.

The numbers of sensitive hosts are 2 for both scenarios.

The optimal path for both scenarios also can be found in scenario files given in the Appendix B. Connection information is also given in the scenario files in detail. The goal of network attack is to compromise sensitive hosts.

1. ***Environments for DQN***

The Reinforcement Learning approach learns the best required policies by interacting with the environment. It is performed by getting certain values randomly, then learning the data required to use for a given state, like Q(s, a), by selecting the task based on current state of policy, implementing the action to environment and updating the values Q(s, a) in state-action relying on the received response as in Figure 2-9. To learn action state values, Q-learning is a policy for temporal differences and it is the reason that Reinforcement Learning Implementations have Q-learning mechanism. It is done by putting the values in the equation (2.1). Here, α is a step size that manages the amount of moving from the current estimate to the new estimate and is a discounting factor which determines how it is shifted from current calculates value to the upcoming results. The calculated future rewards will more accurate and it will specifically determine the optimal values. The quantity of visits to state-action pair is determined with how Q-learning is displaying to meet the optimal action values.

(2.1)

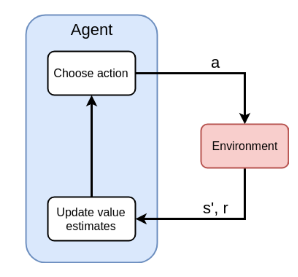


Figure 2-9. Training cycle of Reinforcement learning

RL algorithms are categorized into two main parts according to the determination method of value function Q(s,a).

First is tabular method and second is functional approximation method. Tabular based RL is use a table of state-action to determine value function whereas functional method uses function. Figure 2-10 shows the determination method of value function in tabular method. As we can see from the figure, tabular method uses q-matrix to determine value function from state and action. Tabular method has advantages of easy implementing and high performance to find optimal solution for relatively small problem. However it has disadvantage of bad performance in the relatively large problem since it has to save all state-action pairs in the table. Functional approximation methods, on the other hand, has good performance to the relatively large problem and can generate inexperienced states too, although it is more difficult to implement than tabular method and the performance are mainly affected by selected function. DQN method is a main type of function approximation method and uses deep neural network to approximate function. Figure 2-11 shows the determination method of value function in the DQN method. As we can see from the figure, DQN method uses neural network to determine value function from current state.

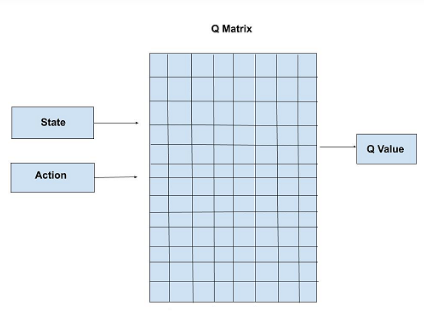


Figure 2-10. Determination of Q-value in tabular method

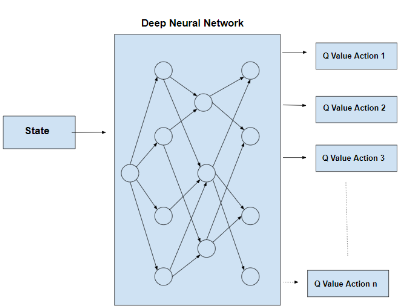


Figure 2-11. Determination of Q-value in DQN

For the high performance of neural network approximating functions, DQN gets high performance to large problems and have been accepting to find solutions of complicating problems.

* NAS: States

State (𝓢) is the set that contains the state of each machine on the network. The state of a machine includes its compromised or uncompromised status, its reachability or unreachability, and the status of its services (present, absent, or unknown). A machine is considered compromised if it has been successfully exploited, and it is considered reachable if it is directly connected to an external network or to a subnet that contains a compromised machine. The state space encompasses all conceivable combinations of service and machine states, including compromised, reachable, and service knowledge.

The state space expands at an exponential rate in proportion to the number of machines and services present on the network. Equation (2.2) illustrates the state space's size, represented by | 𝓢 |, with |E| signifying the number of exploitable services and |M| signifying the number of machines. The exponential base is 3, which indicates that the agent's knowledge of each exploitable service can be either present, absent, or unknown.

(2.2)

**Figure 2-12** shows an example network and its corresponding state. The state reveals that the attacker has successfully taken over the machine at (1, 1) and it is sharing the information with Subnet 2 & Subnet 3 to communicate between the machines, as indicated by the true value of "reachable" for the installed machines in defined subnets above.

Furthermore, the Setup done for the Machine at (3, 1) is elaborated due to action of scanning while the values setup in its configuration for machine 2 at (2, 1) remains are not known. Notably, the state does not incorporate any firewall settings since it necessitates privileged access to determine. Therefore, the simulator assumes that the attacker cannot obtain this information and must learn it indirectly through exploit actions' success or failure. For every service available or accessible on the network, an exploit action for the machine is available and this action is determinable depending on environment chosen by the user and which service is running.

NAS: Actions

In the NAS, the action space consists of a scan action and a deterministic or non-deterministic exploit for each service and machine on the network. The scan action is modeled after the Nmap complete scan and always returns information about service presence or absence. Each exploit action's success is based on factors such as service presence, firewall settings, reachability, and success probability, and each action has an associated cost that can represent metrics such as time, skill, cost, or noise. The reward function in the NAS is defined over a transition (s, a, s'), where s is the starting state, a is the action taken, and s' is the resulting state. The reward for any transition is the amount of any additionally included machine in the future state minus the cost of the action. If no machine is compromised, the reward is simply the cost of the action. The goal of the attacker is to compromise all machines with positive value on the network while minimizing the number or cost of actions used. This aligns with the goals of a real-world attacker looking to gain privileged access or retrieve sensitive information, which can be set in the NAS by assigning certain machines specific values.



If there are any machines comparison, its value is returned in the value ((s’, s) in above formulation from s in the s’ and when none of the machines were comprised, the value is zero and cost of the action a is shown in cost(a).





1. ***Experiments for DQN using NAS and Analysis***

* Deep Q learning algorithms

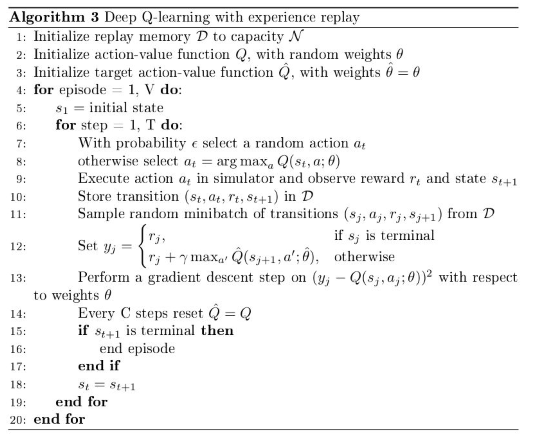
DQN algorithm which is used for penetration testing is shown in figure. Line 1-3 initializes the parameters for DQL. Line 4 and line 6 shows the iteration for episodes and steps. Each episode has several steps of actions.

For each step, algorithm firstly selects the action with epsilon greedy policy and executes it. (Line 7-10)

Then it saves the transition into replay memory D and updates the weights of neural network using gradient descent methods. (Line 11-14)

Finally it terminates if next state is terminal state.

If next state is not a terminal state, it updates current state and go to the next step.



* Steps for Testing.

Tests are devided into two steps.

Firstly we train the DQN agent and check if the training is converged and the goal is reached.

Secondly we test the trained DQN agent using the same environment.

* Parameters for Testing

Main parameters for train are as following.

|  |  |
| --- | --- |
| No | Parameter name |
| 1 | Scenario file name |
| 2 | Observe format |
| 3 | Size of hidden layer |
| 4 | Training steps |
| 5 | Target update frequency |

For tiny scenario, since there are 56 observations and 18 actions, there exist 56 inputs and 18 outputs for neural network. There is no exact method to select the optimal numbers of hidden neurons. The Universal Approximation Theorem states that a neural network with 1 hidden layer can approximate any continuous function for inputs within a specific range. There exists several experienced formula to determine the size of hidden layer. The common principle to set the hidden layer size is that it should be between input layer size and output layer size. We set the numbers of hidden layers as 1 and the number of cells as 32.

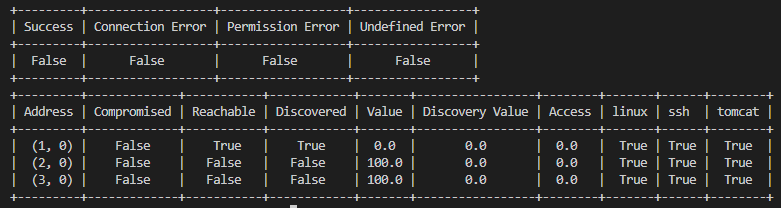
Training steps are set as 10000 and target update frequency is 1000.

Once training ends, we test the DQN agent.

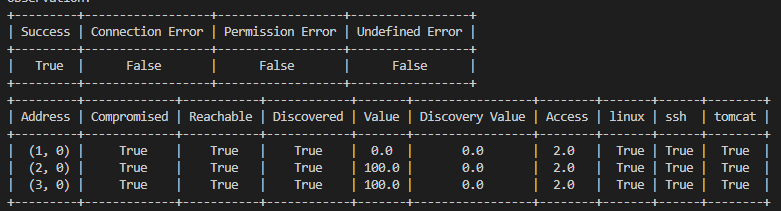
Initial states of networks are given as following figure.

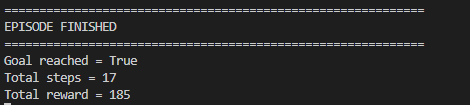
* Rewards

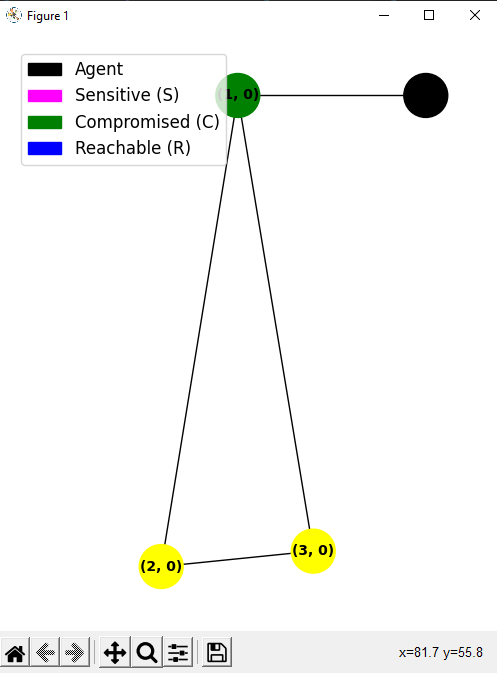
We used the rewards given in the NAS environment.



Final states of networks are given as following.







As we can see from the result, DQN agent find the attack path correctly and reached a goal.

All sensitive nodes are compromized.

Next we test DQN agent for medium scenario.

The structure and initial states of networks in this scenario are as following.

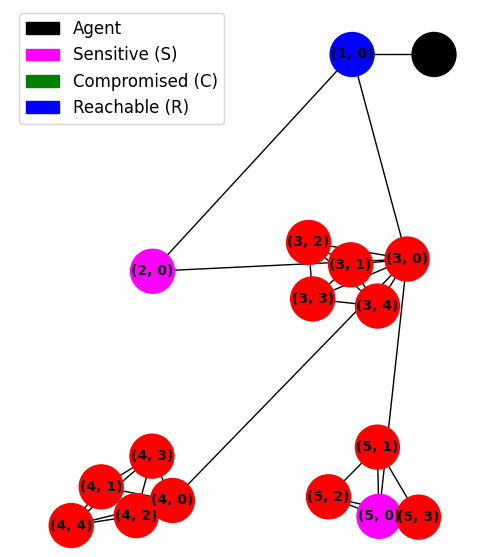
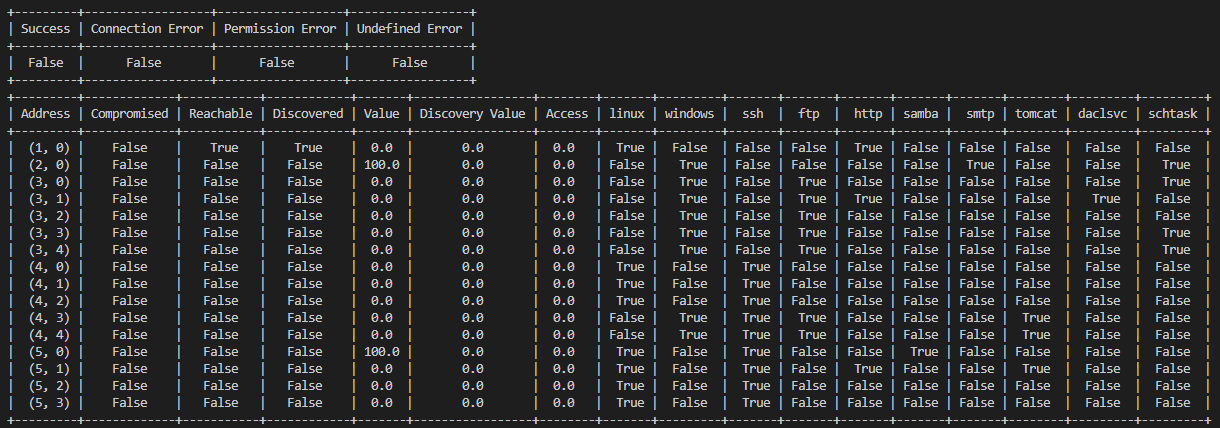


Fig .

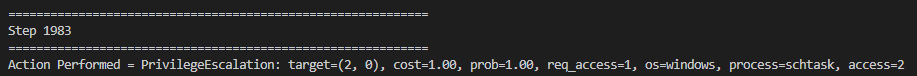


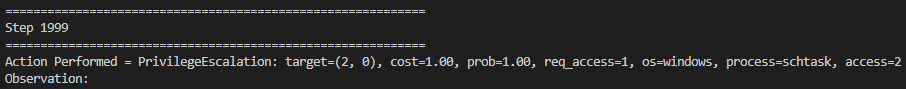
Fig

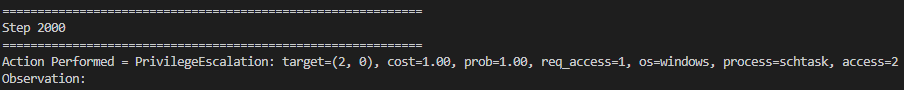
Since the input layer size is 493 and the output layer size is 192, we set the number of hidden layer as 1 and the numbers of cells as 256.

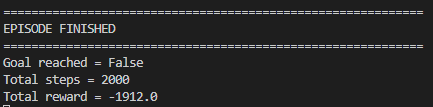
Training steps are set as 1000000.

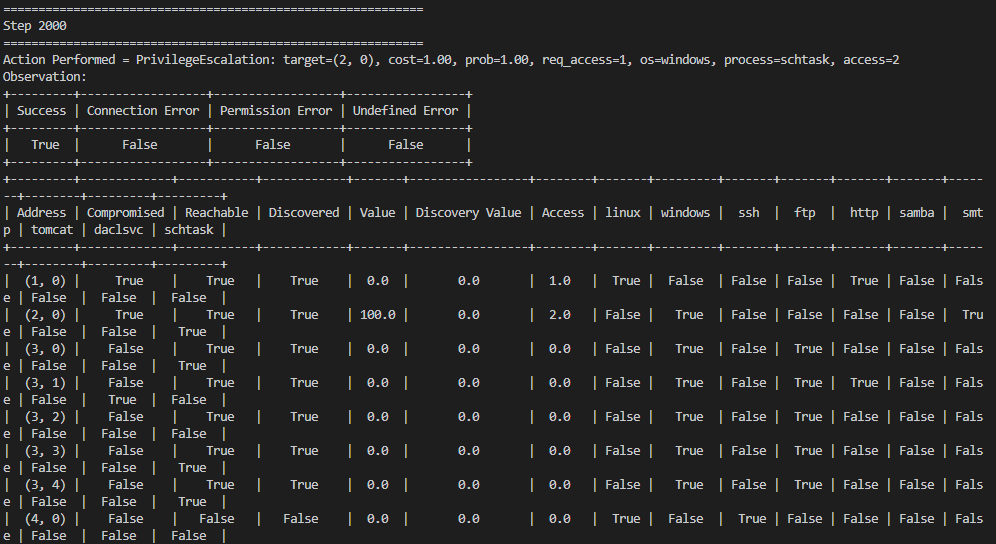
After training ends up, we test the trained model in the same way.

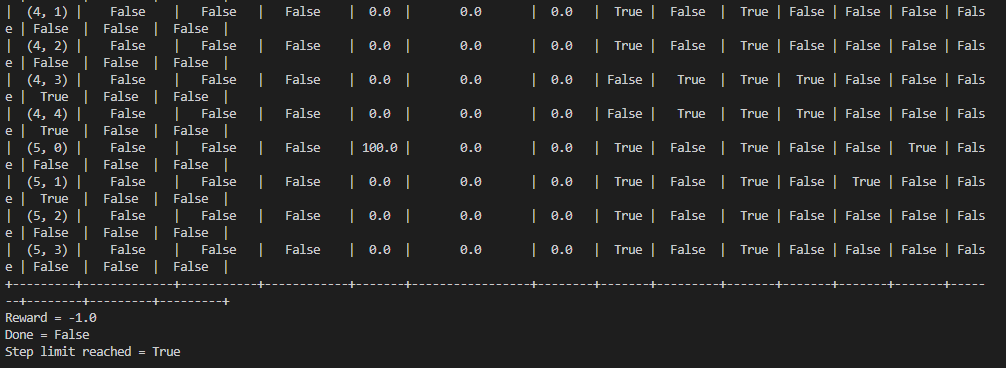


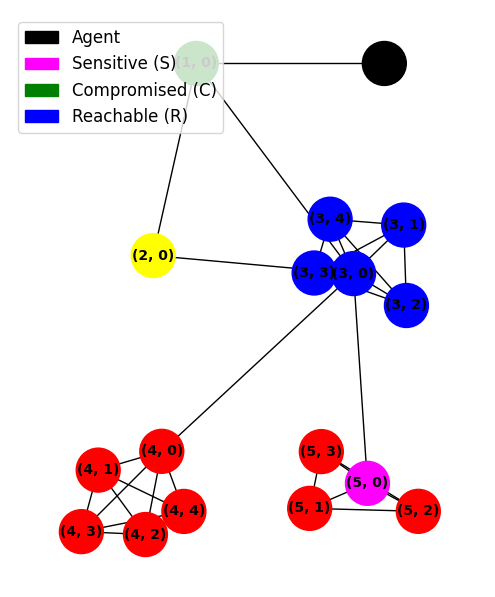












As we can see from the figures, test didn’t reach goals.

From the transitions of states and actions, we can see that agent selected the same action escalation for host {3,0} for last steps.

It means that the agent is put in a infinite loop.

We performed serveral tests on this agent with same environment, however tests showed the similar results and agent didn’t reach the goals.

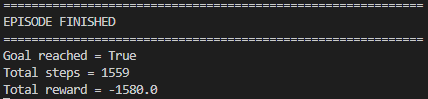
As we showed in algorithm, we use epsilon greedy selction method to select action, however since epsilon greedy is not large, agent gets difficult to avoid infinite loop.

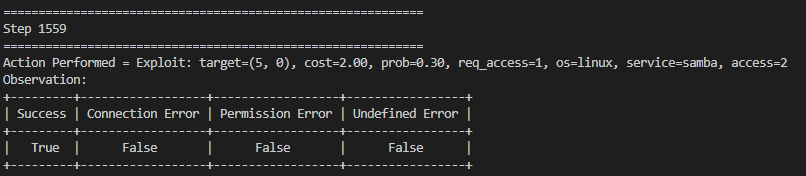
To prevent this, we can set the epsilon as a larger value to select arbitrary action instead of action with maximum q-value. Since we select the arbitrary action with epsilon probability, if we increase it, the probability to select arbitrary action will increase.

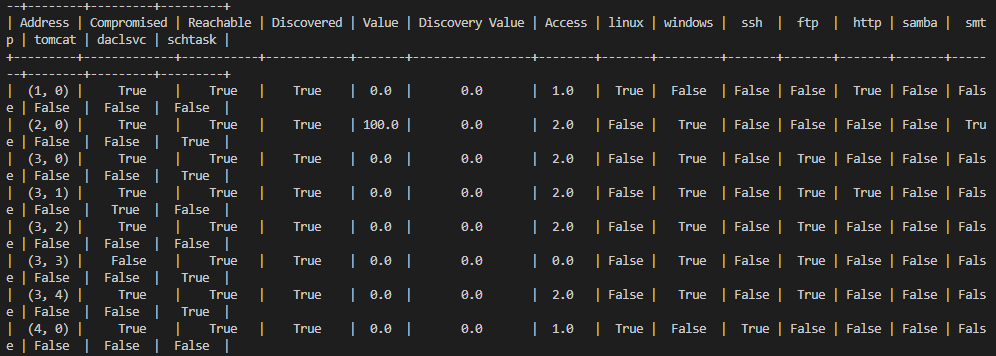
Another issue will be generated if the value of epsilon is set too high as it will downgrade the peformance of the agent. The probability to select a raondom action will be increased and agent’s peformance will be upgrade to avoid infinite loop.

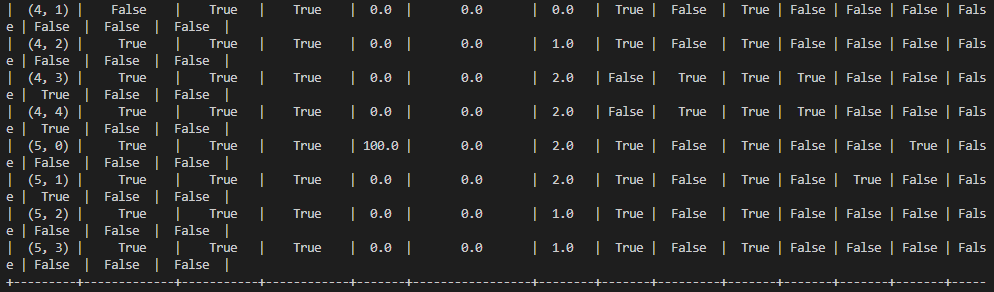
We perform experiment to avoid infinite loop by selecting random agent if the same action is selected repeatedly. We select the epsilon value as 1 if the same action is selected more than three times repeatedly.

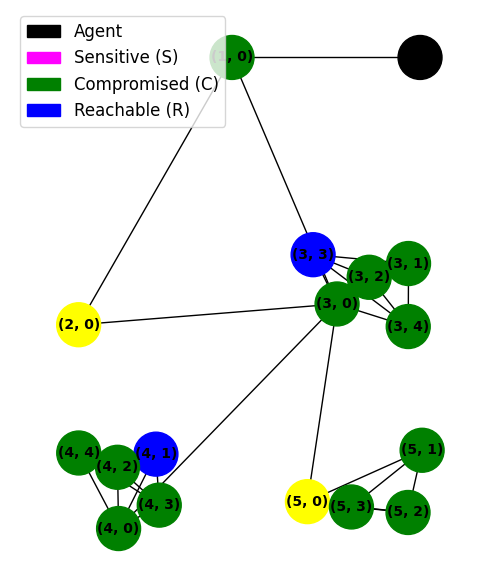
Results for this case are shown in the following.











As we can see from the result, agent avoid from infinite loop and reach a goal.

However we can see that steps to reach a goal gets 1559 steps.

In fact for this scenario, the length and reward of ideal optimal path is 10 and 192. Optimal path is given in the scenario file in Appendix B. We perform experiment for this agent several times and get similar results. As we should perform several experiments in this paper, we are going to set the number for experiments. This is experiment 1.

|  |  |  |
| --- | --- | --- |
|  | Total Steps | Total Reward |
| Experiment 1. | 1371.45+/- 420.41 | -1875.29+/- 660.62 |
| Preceding | 1401.34+/- 317.23 | -1902.32+/-603.47 |

This result is similar as preceding one in [1].

Although a goal is reached, the length of path is too long than optimal path and reward is also too low.

Tests showed that dqn agent in the NAS has good performance for tiny network however has some problems to be solved for big networks.

In the next sub section, we are going to try to analysis the reasons of these problems and find the solution to solve it.

* 1. Analysis

In the previous sub section, we showed the tests for DQN agent in the NAS environment and DQN agent has bad performance for big networks.

As we described in above sections, the numbers of states and actions are increasing rapidly by the numbers of hosts. In fact, net experts who have full knowledge for network attacks can find the attack path for the scenario used in the preceding section since they can find current available attacking methods based on observation.

This means that if we use the information for network attacking methods, we can upgrade the performance of DQN agent largely.

This can be implemented by change reward function of NAS environment. In previous sections, we described for NAS environment.

To improve performance, it is needed to reflect the information of net attacking methods and this can be implemented by updating environment for rewards.

The objective of upgrading is to reduce the length of attacking path. We implemented update for the algorithm in two steps. First step is to study for NAS environment and find the drawbacks of it. For this step, study for reward function seems most important.

Second step is to make updating project based on the result of first step.

Updating for the NAS environment is described in subsection.

1. Updating Performance Auto penetration testing based on DQN agent
   1. **Study on NAS environment**

We described the structure of NAS environment in preceding sections. In this subsection, we are going to study for Reward function of NAS in critically.

Reward of NAS is separated into two parts. First part is value which represents that node was compromised or not compromised and second part is cost which represents the time to do such action.

As we can see easily, second part is to reduce the length of path and find optimal attacking path. This worked well for tiny network however not worked well for the big network as we described in the previous section. First part is used to make DQN agent know the information if the action compromise the host. It makes DQN agent select the action to have high probability to compromise the hosts. It is set as 1 if host is compromised and 0 if not. This term is added to make DQN agent to be converged quickly. However as we described in the previous sections, DQN agent didn’t work correctly as expected. It is because agent didn’t use the information of network attacking methods.

Rewards for current environment are seen as following.

|  |  |
| --- | --- |
| Case | value |
| Action compromise agent | 1 |
| Action didn’t compromise | 0 |

|  |  |
| --- | --- |
| Case | cost |
| Exploit(e\_ssh) | 3 |
| Exploit(e\_ftp) | 1 |
| Exploit(http) | 2 |
| Exploit(e\_samba) | 2 |
| Exploit(e\_smtp) | 3 |
| Privilege escalation | 1 |
| Service scan | 1 |
| OS scan | 1 |
| Subnet scan | 1 |
| Process scan | 1 |

From tables, we can see that reward can only represent the information host is compromised or not but other attacking method information. So DQN agent can’t decide correct action for given environment.

For this reason, training for DQN agent is not performed correctly and second part of reward is not worked correctly. After all, for most cases, rewards are decided by second part mainly and DQN selects the action which has lowest cost for all cases.

This is the reason that DQN agent worked badly for big networks. From analysis for DQN agent, we can know the fact that values for reward should be set considering attacking methods second term for cost can’t work for big network.

* 1. **Updating of DQN agent and NAS environment**
* Study on attack methods

Network attacking has 6 steps of Exploit, privilege escalation, subnet scan, os scan, process scan, and service scan on NAS.

First step of attacking is exploit with several protocols such as ssh, ftp, http, samba and so on.

Once host is exploited, we scan for the subnet which exploited hosts are contained. Privilege escalation is used to attack sensitive hosts and should be performed once exploit to the host is successes. Os scan, process scan and service scan are used to scan protocols and os.

Network experts firstly find the reachable hosts and exploit them. Since network experts are based on information for knowledge, they can’t jump any steps for attacking. For example, network expert will not try privilege escalation for not exploited hosts.

We should reflect this information to the NAS environment to make DQN agent this knowledge.

From the study for current NAS environment and network attacking methods, we propose new reward function for NAS and verify performance through experiments.

From the network steps, we set the rewards as following.

|  |  |
| --- | --- |
| Case | Value |
| Action is failed | 0 |
| Service Scan success | 2 |
| Os scan success | 1 |
| Exploit Success | 4 |
| Privilege Success for general host | 3 |
| Privilege Success for sensitive host | 100 |

As we can see from table, reward value is set reflecting network step.

Exploit is most important step and also first step.

Since success exploit gets greater reward value than others, DQN agent avoids to be put in infinite loop.

Privilege Success for sensitive host gets greatest reward which is much larger than other rewards.

It is since attacks to sensitive hosts are our goal.

It is expected that once we use these rewards, training can get better performance than preceding one.

The problem here is that we can’t consider the cost of attacking correctly so it is difficult to reflect the optimization of attacking path.

We are going to solve this problem by applying cost value as in the preceding one.

However we don’t use the cost value as in the preceding one. Since we set the value based on network attacking methods, there is not mean to set the cost differently for attacking method.

We are going to set the cost as bigger value than the reward value for success attacking steps except escalation to sensitive hosts.

If then, rewards for all cases except success privilege escalation to sensitive hosts will be negative.

Attacking with longer length will have smaller rewards. So we can see that, training for DQN agent will be performed in the direction to reduce the length of attacking path. Though this method can’t reflect the cost of attacking, it will help to reduce the length of attacking path. As we described in previous, cost for current rewards can’t help the DQN agent to find the optimal path for big networks. It can only work correctly for tiny network. For big networks, there are too many actions, DQN agent should consider the network attacking methods to reduce the possible actions and also find the smaller path as possible.

We are going to check the performance of proposed method using experiments for the same scenario.

To see the effects of proposed reward function, we firstly use only first term of reward in the first experiment. Next, we apply the proposed reward to the NAS environment and get the results.

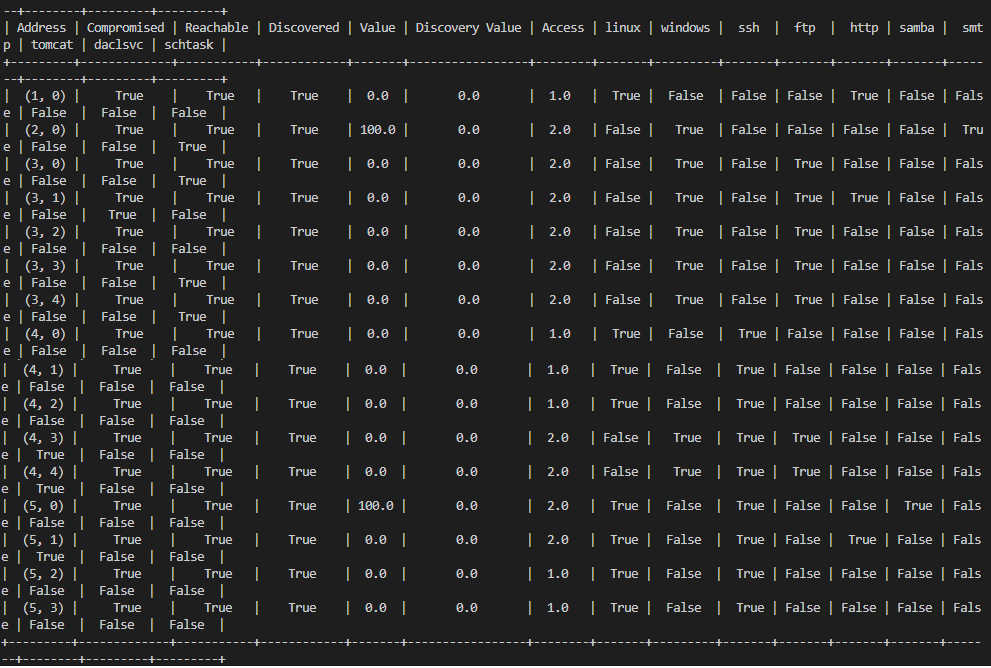
We are going to compare the performance of experiments.

# Experiments / Testing / Measurements / Data

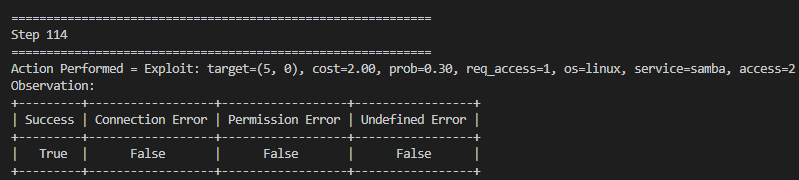
Describe how you have gathered the data for your research

Firstly we perform experiment with the first term of reward. We expected that the training will be success to make DQN agent to reach a goal though it or he will not find an optimal solution. We change the reward value in the NAS environment and test with same scenario file.

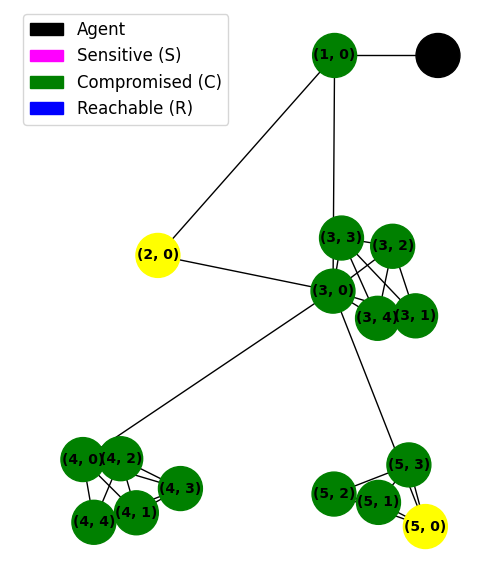
The training steps are set as same as in the previous experiments and other parameters are also same as in the previous experiments. We use proposed reward in the training step only and use the original reward in the test step to compare the performance with original.



I







As we can see from the result, steps are taken 114 and reward is 76. All nodes are compromised in this experiment. It shows that proposed algorithm gets perfect efficient to make DQN agent to find the attack path.

The problem here is that our object is to compromise sensitive hosts as soon as possible. Compromising all hosts to get sensitive information will take much cost.

To analysis the performance of proposed reward, we performed 10 experiments and showed the results in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Total Steps | Total  Reward | Goal Reached |
| 1 | 114 | 76 | Yes |
| 2 | 241 | -53 | Yes |
| 3 | 251 | -61 | Yes |
| 4 | 135 | 57 | Yes |
| 5 | 273 | -83 | Yes |
| 6 | 37 | 155 | Yes |
| 7 | 184 | 12 | Yes |
| 8 | 81 | 109 | Yes |
| 9 | 58 | 128 | Yes |
| 10 | 61 | 127 | Yes |

As we can see from the table, goal is always reached and total steps are in range (37,273) for 10 times experiments.

Rewards are in range (-83 to 127).

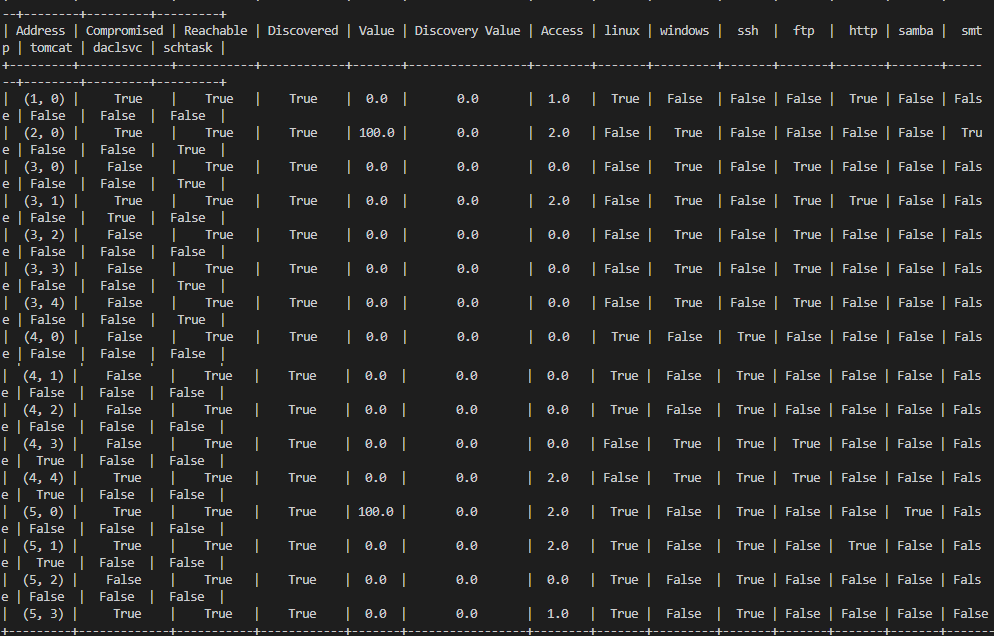
As we described in the previous tests for original algorithm, total steps are reduced 10 times at least and total rewards are increased largely.

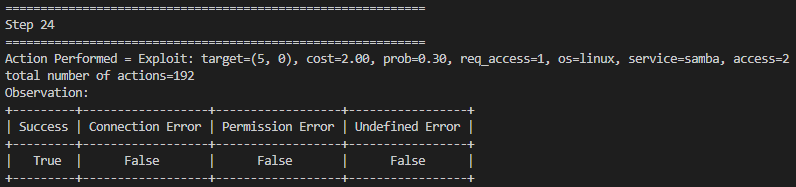
Now, we can go into second step experiment.

In the second experiment we consider the constant cost in reward function and rewards for all actions except for success privilege escalation to sensitive hosts get negative rewards to find small attacking path.

Increment in attacking path length will produce decrement in total rewards and DQN agent will be trained to find the small attacking path.

We perform experiment as the same ways as in first experiment.







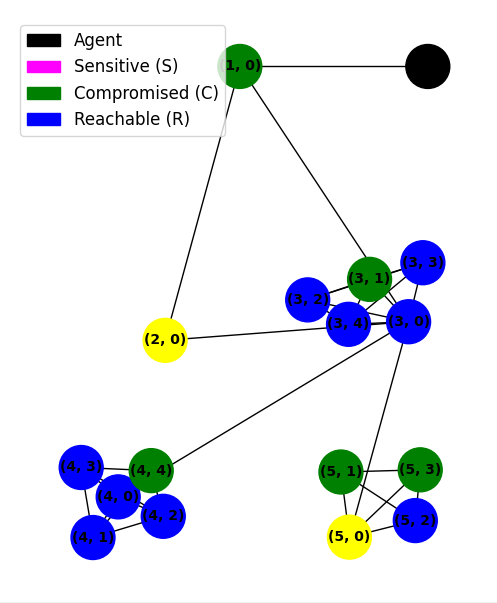
As we can see from the result, total steps are reduced to 24 and reward are increased to 172. This result can’t be get in the previous experiments with maximum value.

Compared to ideal optimal reward 192, it is a very high reaward value. Totol steps are only 24.

For validation, we perform 10 experiments with same condition as same as in the first experiment.

Test resuls are shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| No | Total Steps | Total  Reward | Goal Reached |
| 1 | 24 | 172 | Yes |
| 2 | 35 | 159 | Yes |
| 3 | 61 | 131 | Yes |
| 4 | 43 | 149 | Yes |
| 5 | 18 | 178 | Yes |
| 6 | 144 | 46 | Yes |
| 7 | 130 | 58 | Yes |
| 8 | 22 | 170 | Yes |
| 9 | 100 | 90 | Yes |
| 10 | 57 | 141 | Yes |



# Analysis and Discussion of Results

We are going to analyse the performance of proposed algorithms with experiment results.

We are going to calculate the mean, variation for rewards and steps and compare them with original one.

Following table shows the performance for several methods.

|  |  |  |
| --- | --- | --- |
| Method | Total steps | Total rewards |
| Preceding one | 1371.45+/- 420.41 | -1875.29+/- 660.62 |
| Preceding two | 1401.34+/- 317.23 | -1902.32+/-603.47 |
| Proposed one | 143.500+-27.84 | 46.700+-27.72 |
| Proposed two | 63.400+-14.47 | 129.400+-15.21 |

Results of experiments show that prosed algorithm can upgrade the performance of DQN agent largely.

DQN agent has full ability to find the attacking path with small total steps and high total rewards. A standard network design has been used in this test and a similar architecture was used in previous research studies done for successful testing the system through penetrating testing methods and with available tools in the market.

(EXPLAIN HOW THIS EXPERIMENT IS DIFFERENT THAN PREVIOUS ONES?)

The design used works a simulator of real-world scenario and to test the machine comprised. Also, the generation of random scenarios and test cases with NAS is depending on number of exploits as well as number of the machines included in the simulator testing. In a similar test 3 machines exploited in a 16 machines with two sensitive machines and five services being exploitable.

# Conclusions

In this paper we review for the researches for auto penetration testing based on DQN agent and find the drawbacks. The aim of this study was to know about application of penetrating testing in a simulated environment and to get optimal cases which may incur if a network or machine available on this network are exploited or comprised by an attacker. It is demonstrated that it is possible to utilize RL in simulated environment to determine the system strength and optimally exploiting the network. Through this method, it will be possible to determine which parts of network are including vulnerable points and patching is required in the security of the network. The major advantage of using RL as compared to other Artificial Intelligence Techniques is that RL doesn’t need any previous knowledge of exploits models and a defined standard algorithm to be applied and testing on different networks and network maps, either small setups as in this test simulator or big organization networks.

We proposed new algorithm to update the performance of DQN agent reflecting network attacking methods.

We implemented new algorithm on NAS environment and are extendible to other auto penetration tools.

We concentrate on upgrade of DQN agent to work well for big networks and find the short attacking path.

From experiments we validated that proposed DQN agent has ability to work well instead network experts for big networks.

(EXPLAIN IF IT IS SCALABLE?)

# Future Work

In this research, we can see that applying Reinforcement learning methods to auto penetration testing, environment has a very important role.

We concentrate on update of Reward function in NAS.

If we consider the network attacking methods to the reward function, we can upgrade the performances of DQN agents of other auto penetration tools.

We are going to study for to reflect network attacking methods to other penetration testing tools which use RL methods.

Appendix A: Literature Review

{4} This work is long rooted and in line with parts of applied research focused on automating and optimizing offensive cyber security auditing processes, with a specific emphasis on vulnerability assessment (VA) and penetration testing (PT) [2,11on mo]. In this context, the authors provide a summary of previous research efforts, highlighting the adopted approaches and significant contributions.

Initially, researchers were focused on the planning phase of these processes, with some works implemented in industrial PT systems and frameworks, while others remained in the realm of research ideas [2,4]. As the domain of PT automation and enhancement lies at the intersection of cyber security and AI research fields, multiple avenues of research have been pursued through different methodologies of automated planning (a sub-area of AI) [6,7].

Early research on this topic was about modelling penetration and how penetrating testing is used in decision making approach [8]. Although most of the works were initially focused on vulnerability assessment, they paved the way for significant contributions to the PT domain as well.

To provide a comprehensive understanding of the research in this field, the paper presents a literature review section summarizing previous research efforts, with a focus on the adopted approaches and contributions. To get an optimal result, the authors today present the actual results of the search in this field and then further divide it for research in approach, methodology and by the type. [6].

1. **A detail of previous works done on Penetrating Testing automation**

When the object of a system is to develop efficient PT system for the organization, the most famous approach for PT tasks is the automation.

If it is don’t in an appropriate way, we may fail to achieve the goals as each phase of automated testing for main tasks and their sub-tasks or activity testing is highly challenging, e.g. the utilization of automating systems and tools that may be performing their tasks, their actions and the results without any pre-processing or optimization may be a huge failure to get the optimal results. The automated systems may not be able to produce the required results in small or large environments if there is no permanent control of a PT human expert reviewing the process. It is because automated PT require large number of a complete network structure[8–11]. In Penetrating testing environments, there are a number of issues which incur during the process in addition to actual time duration of the tests. The issues which are highlighted to automation are a big number of false positive alerts notified on the assets such as firewalls and IDPS and network congestion or generated traffic. Therefore automated Penetrating Testing was only limited to networks of small size or medium size and those with customized scripts to perform only in Penetrating testing.

Initially, research focused on improving the PT system by enhancing the planning phase, which was approached through attack graphs or decision trees, reflecting the PT practice's nature as sequential decision-making. However, most of the works focused on vulnerabilities assessment rather than PT, as the approach was static and limited to the planning phase. One notable contribution was the modelling of VA as atomic components, pre-conditions, and post-conditions to narrow down the targeted vulnerability. Later works attempted to automate the planning phase of PT tasks, but this approach did not address the issue of improving performance and only focused on the planning phase.

Notwithstanding, one remarkable work introduced optimization by modelling PT as planning domain definition language, which calculated the Penetrating Testing Post-attacking and attacking phases, allowing for integration with some PT systems. This solution generated different types of attack plans for single and multi-paths PT scenarios and was executed in conjunction with information-gathering tools, which transformed the acquired information into input for a planning problem to be solved separately. The proposed approach was limited to small and medium-size networks, however.

Some research considered AI to enhance PT practice, but most of the proposed models were not successful to cope with continuous uncertainty in Penetrating Testing practices and unavailability of reliable and accurate knowledge about the systems assessed. A possible way was to use the ML algorithms in VA System and skilled Penetrating testing called Core-Impact, which modelled the PT planning phase as a partially observable Markov decision process resolved with a solver from external POMDP to calculate the best testing plan in the form of attack vectors. However, this model was insufficient to include the pending tasks and testing phases, such as pivoting, testing and vulnerability assessment that should be non-standard, sequential and highly interactive when matched with information collection start from planning phases.

1. **Major Drawbacks and Limitations of Current Penetrating Testing Practice**

We will present the current Penetrating Testing Practices, tools, systems and frameworks, limitations and major drawbacks of these practices along with domain of the penetrating testing and automating PT Practices. On Large networks, the penetrating testing becomes slow because it includes repetitive and routine tasks. Although much of the penetrating testing routine can be automated, the repetitive tasks are crucial and can’t be skipped resulting in slowing down on large networks [6]. While they were in theory and in the documents, many of proposed solutions were looking affective and to help solve the problem, the practical demonstration of penetrating testing proved that added improvement were not good enough to solve the core issues along with available resources and infrastructure.

On the other hand, some solutions were not adequate in the context of penetrating testings. It is evident today that human computing capabilities are limited and time consuming as compared to machines when it comes to today complex tasks and large algorithms. It takes days or weeks to an average Penetrating tester while working on a medium sized Local Area Network and core focus is that testing on a specified network should be comprehensive and cover all available services and machines on that network. Along allocate time and efforts, a large amount of system downtime will be reported while performing the tasks. In terms of accuracy and quality of the results, the above two points add to the poor performance and it may be a fact that human beings change their opinion, get bored and commits mistakes.

In response to the above issues, **Penetrating Testing Automations** were presented as a perfect solution. Therefore, semi-automated or fully automated solutions were designed with their main goal to reduce the frequencies of testing, testing coverage increase and save time and reduce human labour. In terms of defined approach, the proposed solution was very affective and adoptive to rely on automated planning generates an automated attack plan. This attack plan was named as attack graph and then functioning their tasks in automated ways. Meanwhile, the newly added system solutions worked to perform complex tasks and shorten the overall process along with adding new cases of exploits.

The Existing Penetrating Testing tools, frameworks and systems could be easily automated and therefore the limits of these existing PT System were the major questions of Cyber threats security community. These PT Systems can perform majority of PT Testing tasks with little or no human involvement, especially during discovery of the Vulnerability and information gathering- the first two phases of Penetrating Testing.

The companies which have constant need of their security audit, the fully automated Penetrating Testing Systems which may perform basic and advanced testing without any human presence are more demanding to them. In spite of that, the researcher had been in continuous struggles to find Penetrating Testing Tools which may automatically perform their sophisticated procedures as human can hardly master the processes due to extent of their complexity. Therefore, developing a machine that would replace the PT experts is a major challenge and interest of the researchers and also because PT practise has multi-phase nature with dependency on task and phases of the testing.

The process of performing a PT (Penetration Testing) can be complex, and one major issue is the incomplete profile of the assessed system that is produced during gathering the information and reconnaissance phase of the PT.

Experts often deal with this issue by repeating tasks, changing their approach, or making assumptions to continue tests. Additionally, modern attacks have adopted evasive techniques and complex paths, allowing them to bypass network and system defences. Skilled attackers exploit a chain of vulnerabilities, including hidden and composite vulnerabilities, in networks, making it difficult to secure the entire infrastructure.

The capacity to find and evaluate the interactivity of several possible vulnerabilities on a system or network guide to a better learning of whole set of vulnerability of the assessed system. At the end, the penetrating Testing Data output is important, but it is often not utilized at the time of future tasks or retesting and it is put at a side after generation of PT Report.

Our research aims to address this issue by focusing on using previous test results when retesting is required, as many of the system architectures and security configurations do not change remarkably in a short or long term.

1. **Motivation and Contribution**

The computer systems are not exception to the rule of High Complexities threating the controls and security of the systems. It is general understanding that increase complexity becomes an threat to the existing Ecosystem of Security and Controls and same applies to the computer systems. The Security professional avoided the trap of bolting policies and security layers; a classic approach of defending digital assets and networks from external cyber security threats and they focused their thoughts on offensive security.

The Security Professional were defied with the requirements of an Intelligent Penetrating Testing framework and System which should help manual human testing work. There was much focus on complexity of the testing systems and high demand of PT to perform more or less of automated tasks, while taking it over from human interference and conduct data collection, assessment of vulnerabilities and risk factors, exploits with more time to Security Professional to research and work on complex solutions of Post-Exploiting results.

Based on our mentioned facts about Penetrating Testing practices for evaluating the systems

Security, Machine Learning (ML) is found to be the best solution of that problem. In initial phases, many Artificial Intelligence (AI) techniques were introduced, but Machine Learning was most sophisticated approach selected for automated PT system and act like a professional security Tester learning and gaining skills over the time and implementing it in real time testing scenarios.

This research will highlight gap between current Penetrating Testing practice and focus on solving following key issues:

* To Reduce Testing system cost due to Human labour cost and moving towards systematic testing as an alternative to it.
* To Allow adaptability and flexibility while dealing more affectivity with different types of known security threats.
* Significantly mitigate security risks, improve performance, and minimize downtime during testing to minimize the impact on the assessed network.•
* To free Testing Experts form repetitive and boring testing tasks and assign them on more challenging and purposeful assignments.
* To perform a wide variety of security exploits and makes use of evasive and complex paths of the network which may be hard to find if investigated manually by the human.

We may summarize it that cyber attackers aim to get specific access through exploitations of the installed systems by going through a chain of attacks and to find any possible vulnerability.

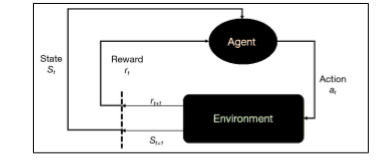
The objective of IAPTS is to automatically detect, utilize, and manage network machines and networking devices, while uncovering concealed and intricate security flaws within network infrastructure or segments. In essence, IAPTS should encompass all phases of penetration testing, going behind the vulnerability assessment offered so far.

The systems which prove to be secure while working in an individual control, it also helps to cover the testing and secure the locally secured systems but an addition of both may result open a loophole for the hacker to exploit through it. At the end, the Penetrating Testing expertise re-utilization, generalization and extraction is the concluding reward of IAPTS to the prevailing Penetrating Testing Practices.

The realm of computer science and artificial intelligence (AI) has witnessed a significant increase in research related to the creation and evaluation of intelligent decision-making systems. IAPTS is a system that perceives its environment and independently determines the most efficient approach to conduct PT tasks. Cyber security solutions powered by AI typically fall under two categories: expert-led and automated systems that use unsupervised machine learning. The former, such as AVs, FWs, IDPSs, and SIEMs, rely on input from security professionals and may be prone to errors.

1. **Reinforcement learning (RL)** techniques guide the existence of autonomous or semi-autonomous decision-making systems that reflect real-world contexts, particularly in the offensive security domain, such as vulnerabilities assessment and PT context. RL was selected for its capacity for autonomous learning, reward-based learning, and the richness of its environment, which can capture all major features of PT, including uncertainty and complexity. Reinforcement learning is a branch of AI that enables software agents to determine optimal behaviour within a given context to maximize performance. The agent adjusts its behaviour based on simple feedback (reward) from the environment, learning and adapting to improve its decision policy, if necessary. A best approach to maximize the overall reward by some Reinforcement Learning Algorithms which can cover and optimize to global requirements.

The RL environment is explored, and the agent learns how to act based on the rewards received from its actions, as shown in Figure 2.



**Figure2.** In this figure, the state of Environment **x(t)** is being observed by the Reinforcement Learning Agent at a time **t** and on the basis of its Policy for selective active, it selects the action **a(t). Then it get a reward r(t+1)** after transitioning at time **( t+1 )**to the state **x(t+1).** As the agent gains experience, it becomes increasingly adept at selecting actions that result in the most significant combined reward.

The Human Intervetion is significantlly reduced or totally removed with use of Reinformcement Learning method. Furtheremore, the time required to learn and customize is less as it is the case with expert systems and machine learning corelatively. The suitibility of Reinformcement Learning for increasing the automation of Penetrating Testing solutions has been greatly testing as said above and many new algorithisms have been tested and added. These new algoritims are implemente along with new tool boxes and an enchanced usibility to answer complex problems of RL in restricted resources scenarios with much affective results.

* 1. **Towards Partially Observable Markov Decision Making Process(POMDP)**

Partially Observable Markov Decision Process(POMDP) is an another approach that involves modelling an attack on a computer system with attacker incomplete knowledge in the simulation environment with configuration and learning to threats continue as the threats are applied on the system. It is affective when applying a Penetration Testing on a single host Machine.

The concept of Penetration Testing (PT) involves a series of tasks that can be carried out manually or by an automated PT platform to achieve a predetermined or unknown goal, which is referred to as the target. The target could be a computer system, network, or information stored on a computer. In some cases, the attack may not have a specific target. The goal of this study is to create an efficient and optimized autonomous PT system that utilizes Reinforcement Learning (RL) and other techniques to enhance its performance, efficiency, testing coverage, and reliability. To model PT as an RL problem, the researchers employed Partially Observable Markov Decision Process (POMDP). The POMDP model comprises a tuple M=<S,A,O,T,Ω,R,b1> with finite states, actions, and observations in S, A, and O, respectively. The agent interacts with the environment by executing an action, receiving an observation and a reward, without directly obtaining information about the state of the environment. The choice of using POMDP for PT modelling was previously explained in a publication, and some of the key points are summarized in the following sections. Since an attacker would have incomplete knowledge and not sure about which of the hack scans may succeed while actually performing in on the system, the POMDP approach helps in defining the most relevant properties of real world hacking.

The use of Partially Observable Markov decision processes (POMDPs) has proven to be a powerful tool in solving planning problems that involve hidden state and uncertainty in action effects. Recently, algorithms That employ an approximate value iteration technique with point-based update updates have been developed, which have been approved to measure correctly.

These algorithms depend on executing several highly quick approx. updates over a set of beliefs drawn from the reachable part of the belief simplex.

By calculating a bound on the estimated error that is proportional to the sample spacing of the set, one can tolerate more calculation error at the places that can only be reached after many time steps in discounted problems. This idea is based on a convergence result that is derived from the fact that the set samples enough of the search tree to some depth. However, fully expanding the search tree to depth t requires a number of points exponential in t.

This article introduces a novel convergence argument that combines elements from both approaches, and applies to cases where the varying of sample spaces is according to what is known as discounted reachability. By more accurately reflecting the behaviour of current algorithms, this new approach is expected to provide improved performance and efficiency.

**Partially Observable Markov Decision Process (POMDP) Solving Algorithm**

There are complex and sequence of decisions in Reinforcements Learnings(RL) methods while solving real- world problems in POMDPs and in the form of MDPs. In these models the available opportunities later affect each of the decision made.

However, we are not concerned about improvement or development of new Reinforcement Learning algorithms or methods in this work, rather to search and apply the right algorithm that would be pertinent to find appropriate results and a precise solution for our problems.

We may look for an appropriate choice of approach and resolving the algorithm to be made because Reinforcement Learnings Methods are usually completed when solving the large and complex problems. Instead of Selecting One algorithm, the IAPTS should depend on different algorithms when solving the complex environment of Penetrating Testing Partially Observable Markov Decision Process (POMDP). It will be become simpler and more adaptable that IAPTS should select different algorisms for solving the complex problems in POMDP environments in Penetration testing. To work according to context or available scenario, the IAPTS will change itself to select and use most appropriate approach of solving the problem. Moreover , the use of one algorithm for problem solving may make challenging for IAPTS in terms of available Computational resources, Memory and Time resources, so the choice of selecting different algorithms is judgemental for IAPTS and selecting a right approach by which accuracy is usually immolated to acceptance.

While working with large number of transactions or choosing for a scheme of static reward, it is important to note that large environments will also create the big challenges in solving the algorithms. There two main categories of Solving Algorithms used in Reinforcements learnings, including

1. Policy Search Solving
2. Oriented Solving

These are the two different approaches to solving the problem of maximizing rewards in a reinforcement learning (RL) environment. The first approach, known as the reward approach, involves developing an optimized rewarding function that takes into account both immediate and long-term rewards. However, this approach can be time-consuming and complex if the problem representation is not optimized.

The second approach, policy search, focuses on constructing a decision policy graph that maximizes the long-term reward. To implement both approaches, the researchers used both on-policy and off-policy implementation methods, and used a perfect already designed POMDP-solving approach to evaluate the performance of different solving algorithms. The article concludes by mentioning the shortlisted algorithms that were initially considered.

The reward approach involves developing a comprehensive and optimized rewarding function that considers both immediate and long-term rewards to determine the best possible rewarding scheme for each transition and observation. However, this approach can be time-consuming and complex, especially if the problem representation is not optimized.

The decision policy graph is constructed by policy search in the other way around and it is done by understanding the policies that would impact and maximize long term reward and do internal environment internal mapping.

In this research, both reward-optimization and policy-search approaches were used, along with on-policy and off-policy implementation methods to ensure the best quality policies were found. To evaluate the performance of different solving algorithms, the researchers used a powerful off-the-shelf POMDP-solver and state-of-the-art algorithms, excluding external factors. Several algorithms were initially shortlisted for consideration.

* 1. **PERSEUS Algorithm**

The PERSEUS Algorithm is a ground-breaking approach to solving partially observable Markov decision processes (POMDPs) that offers a more efficient and effective solution to the problem of agent planning under uncertainty. Unlike other point-based methods, PERSEUS performs approximate value backup stages that improve the value of many belief points with a single backup. By selecting a randomly chosen subset of points within the belief set, the algorithm can back up only those points sufficient for improving the value of each belief point in the set. Additionally, PERSEUS perform a policy-iteration along with a single back when it searches through finite state for stochastic space.

This approach makes PERSEUS an exceptionally efficient algorithm for solving large scale POMDP problems. With its approximate solving nature and the ability to work and perform on a surplus theory set illustrated by simulating decision sequences, PERSEUS shows tremendous potential for accelerating the solving process and achieving optimal results. The algorithm is a breakthrough in POMDP research and offers a new path forward for agent planning under uncertainty.

* 1. **GIP Algorithm**

The GIP algorithm is revolutionizing POMDP exact solving methods with its reliance on incremental pruning. Instead of using LPs to check for dominating vectors, GIP uses a Benders decomposition and only a fraction of the original LP constraints. In fact, GIP has been proven to outperform common vector pruning algorithms for POMDPs, reducing running time and memory usage in large POMDP contexts. The latest version of GIP is considered the fastest optimal pruning-based POMDP available.

In exciting news, the GIP algorithm has been updated to allow for external belief sampling, eliminating the need to sample beliefs from the POMDP environment at the beginning of the solving process. This change will enable agents to upload their beliefs directly to the RL environment for more efficient use. Overall, GIP is a game-changer in the world of POMDP solving algorithms, and its latest updates make it even more effective and accessible for users.

The GIP (Generalized Iterative Policy) algorithm is a reinforcement learning algorithm that can be used in Partially Observable Markov Decision Processes (POMDPs). POMDPs are a type of decision-making problem where the agent does not have access to the full state of the environment, but only observes a partial state through sensors. The goal of the agent is to take actions that maximize a long-term reward.

In POMDPs, the optimal policy is a mapping from the current belief state (a probability distribution over the possible states) to actions. The GIP algorithm can be used to learn the optimal policy in POMDPs by iteratively improving an initial policy. The GIP algorithm works by first computing the value function for the current policy. The value function represents the expected discounted reward for each belief state under the current policy. This can be computed using dynamic programming techniques such as the Bellman equation.

Once the value function has been computed, the GIP algorithm updates the policy by greedily selecting the action that maximizes the expected reward for the current belief state, based on the value function. This process is repeated iteratively until convergence.

The GIP algorithm has been shown to be effective in solving POMDPs in a variety of domains, including robotics, autonomous driving, and game playing. However, it can be computationally expensive and may not scale well to large state spaces. Therefore, researchers continue to explore new algorithms and techniques to address these challenges.

* 1. **PEGASUS Algorithm**

PEGASUS is a cutting-edge policy-search algorithm designed to tackle the challenge of solving large Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs). This innovative approach, initially proposed by [20], involves transforming any MDP or POMDP into an equivalent POMDP in which all state transitions are deterministic. It decreases the impact of problem for policy search to the one where consideration is done with transactions with POMDPs only.

To further simplify the policy search process, PEGASUS estimates the value of all policies, allowing for easy identification of high-value policies. This purposeful algorithm has been presented already its capabilities in solving complex problems, because it generates a polynomial dependencies on horizontal time rather than exponential dependencies. As a result, PEGASUS is an ideal solution for penetration testing POMDP solving, and has paved the way for even more advanced policy-search algorithms in the future.

Typically, policy search methods require access to a POMDP either by executing trajectories or using a black-box "generative model" that enables the learner to try actions from arbitrary states. In this paper, however, we assume a more powerful model than these - an implementation of a generative model without an internal random number generator. This means the model has to ask us for random numbers whenever it needs them, such as when it needs a source of randomness to draw samples from the POMDP's transition distributions. This simple change results in what we call a deterministic simulative model, which turns out to be surprisingly powerful.

Using this deterministic simulative model, we demonstrate how we can reduce the problem of policy search in any arbitrary POMDP to one in which all the transitions are deterministic. This means that taking an action in a state will always deterministically result in transitioning to some fixed state, except for the initial state which may still be random. This reduction is achieved by changing/reshaping the actual POMDP into a similar one with only calculating transitions.

The policy search algorithm we selected then works on this transformed POMDP, searching for a high-value policy. By assuming access to a deterministic simulative model, we are able to significantly simplify the policy search problem, which has implications for a range of applications in which POMDPs are used.

* 1. **Other Candidates**

Some other Reinforcement Algorithms like Finite Grid and Backwards Induction along with addition of the candidates. These RL Algorithms are used to find a shortest path when more than one policy is available and it is a point-based value iteration. POMDP-solver software have already many of proposed algorithms included in it, and a emphasized implementation is provided along with its power to improve its version over the time.

"Revolutionary algorithm implemented to bridge the gap between theoretical research and real-world industry practices. Highly recommended 'ready solution' utilized for benchmarking purposes, yielding promising results. Ground-breaking Palm leaf search (PLEASE) algorithm evaluated and proven effective in solving complex POMDPs with large observation spaces." [24]. In Reinforcement Learning, a finite grid is a common setting for many problems. A finite grid is a graph of nodes or states where each state represents a specific location in the grid. These states can be thought of as discrete points in space, and an agent in a particular state can take actions to move to other adjacent states. The objective of the agent is to learn a policy that maximizes its long-term rewards in this grid environment. Backward Induction is a common algorithm used in Reinforcement Learning to solve finite grid problems. It is a dynamic programming method that works by starting from the last state and calculating the optimal value of each state recursively. The optimal value of a state is the expected sum of rewards that the agent can accumulate by following the optimal policy from that state. The backward induction algorithm starts by initializing the value of the last state in the grid to zero, and then it moves backward to the previous state. For each state, the algorithm calculates the optimal value by considering all possible actions that the agent can take from that state and the resulting values of the next states. The optimal value is then stored and used to calculate the optimal value of the previous state. The process continues until the optimal values of all states in the grid are calculated.

Once the optimal values of all states are known, the backward induction algorithm can calculate the optimal policy for the agent. The optimal policy is the set of actions that the agent should take at each state to maximize its long-term rewards. The policy can be determined by selecting the action that leads to the state with the highest optimal value.

In conclusion, the finite grid and backward induction algorithms are essential tools in Reinforcement Learning. They enable agents to learn optimal policies in finite grid environments and can be extended to more complex settings with larger state spaces. These algorithms have practical applications in areas such as robotics, gaming, and finance, where agents need to make optimal decisions in uncertain and dynamic environments.

* 1. **Deep Reinforcement Learning Algorithms**

In recent years, there has been a surge of interest in Deep Reinforcement Learning (DRL) algorithms, which have demonstrated impressive results across various domains. These algorithms operate similarly to traditional reinforcement learning methods, but they incorporate deep neural networks for greater adaptability to environmental changes. One exciting application of DRL is in the field of auto penetration testing, where it has shown promising results.

Researchers have proposed DRL-based auto penetration testing methods using the NAS and Python Tensorflow packages, implementing deep Q-learning agents for testing in NAS environments. While the results in smaller scenarios were encouraging, larger networks have yet to be fully explored.

Another proposed method involved the A3C model, which focused on reducing training time in DRL-based penetration testing by utilizing parallel computation. This approach has shown promise in reducing the overall time required for testing and could be acceptable for post-penetration testing scenarios.

Overall, DRL algorithms offer flexibility, efficiency, and scalability, making them a promising tool for a variety of applications, including auto penetration testing. While more research is needed to fully realize their potential, DRL algorithms are certainly at the forefront of revolutionizing machine learning.

DRL algorithms build on this concept by utilizing deep neural networks to represent the agent's policy, value function, or both. These deep networks can handle high-dimensional inputs, making them suitable for solving problems with complex state spaces, such as image or video data. DRL algorithms have proven to be particularly effective in solving problems where traditional techniques have failed, such as executing complex penetrating testing scenarios. A major advantage advantages of DRL algorithms is their ability to learn from raw sensory inputs. This means that the agent can learn to perceive and understand the environment without any pre-processing or feature extraction. For example, in the context of robotics, DRL algorithms can learn to control robotic arms by directly observing the environment through cameras, without requiring any manual calibration or feature engineering.

Another important feature of DRL algorithms is their ability to learn from sparse rewards. In many real-world problems, the reward signal is only available intermittently or in a delayed manner, making it difficult for traditional techniques to learn. DRL algorithms can learn from such sparse rewards by leveraging the temporal structure of the problem, i.e., by estimating the value of a state or action based on the expected future rewards.

DRL algorithms can be broadly categorized into two types: model-based and model-free. Model-based DRL algorithms involve learning a model of the environment, i.e., a function that predicts the next state and reward given the current state and action. Model-based algorithms can be effective in situations where the environment is well-understood and can be accurately modelled. However, they can be computationally expensive and may not scale well to large or complex environments.

Model-free DRL algorithms, on the other hand, do not require a model of the environment. Instead, they learn the policy or value function directly from the raw sensory inputs. Model-free algorithms can be more efficient and scalable than model-based algorithms, but they may require more data to achieve good performance.

Some of the most popular DRL algorithms include Q-learning, SARSA, Deep Q-Networks (DQNs), Actor-Critic, and Proximal Policy Optimization (PPO). Q-learning and SARSA are traditional reinforcement learning algorithms that can be extended to deep neural networks. DQNs were one of the first DRL algorithms to achieve human-level performance in video games. Actor-Critic algorithms combine the advantages of policy-based and value-based methods. PPO is a state-of-the-art algorithm that uses a trust region optimization approach to improve stability and convergence.

In conclusion, Deep Reinforcement Learning algorithms have emerged as a powerful tool for solving complex problems in various domains. By combining the concept of reinforcement learning with deep neural networks, DRL algorithms can handle high-dimensional inputs, learn from raw sensory inputs, and learn from sparse rewards. While DRL algorithms have achieved impressive results in recent years, there are still many challenges and opportunities for future research, such as improving sample efficiency, developing better exploration strategies, and adapting to changing environments.

Appendix B: Data and Programs Used for this Report[[5]](#footnote-5)

*(required section and optional section)*

1. Scenario files

- Tiny

# A tiny standard (one public network) network configuration

#

# 3 hosts

# 3 subnets

# 1 service

# 1 process

# 1 os

# 1 exploit

# 1 privilege escalation

#

# Optimal path:

# (e\_ssh, (1, 0)) -> subnet\_scan -> (e\_ssh, (3, 0)) -> (pe\_tomcat, (3, 0))

#     -> (e\_ssh, (2, 0)) -> (pe\_tomcat, (2, 0))

# Score = 200 - (6\*1) = 194

#

subnets: [1, 1, 1]

topology: [[ 1, 1, 0, 0],

           [ 1, 1, 1, 1],

           [ 0, 1, 1, 1],

           [ 0, 1, 1, 1]]

sensitive\_hosts:

  (2, 0): 100

  (3, 0): 100

os:

  - linux

services:

  - ssh

processes:

  - tomcat

exploits:

  e\_ssh:

    service: ssh

    os: linux

    prob: 0.8

    cost: 1

    access: user

privilege\_escalation:

  pe\_tomcat:

    process: tomcat

    os: linux

    prob: 1.0

    cost: 1

    access: root

service\_scan\_cost: 1

os\_scan\_cost: 1

subnet\_scan\_cost: 1

process\_scan\_cost: 1

host\_configurations:

  (1, 0):

    os: linux

    services: [ssh]

    processes: [tomcat]

    # which services to deny between individual hosts

    firewall:

      (3, 0): [ssh]

  (2, 0):

    os: linux

    services: [ssh]

    processes: [tomcat]

    firewall:

      (1, 0): [ssh]

  (3, 0):

    os: linux

    services: [ssh]

    processes: [tomcat]

# two row for each connection between subnets as defined by topology

# one for each direction of connection

# list which services to allow

firewall:

  (0, 1): [ssh]

  (1, 0): []

  (1, 2): []

  (2, 1): [ssh]

  (1, 3): [ssh]

  (3, 1): [ssh]

  (2, 3): [ssh]

  (3, 2): [ssh]

step\_limit: 1000

* Medium

# A medium standard (one public subnet) network configuration

#

# 16 hosts

# 5 subnets

# 2 OS

# 5 services

# 3 processes

# 5 exploits

# 3 priv esc

#

# |A| = 16 \* (5 + 3 + 4) = 192

#

# Optimal path:

#  (e\_http, (1, 0)) -> subnet\_scan -> (e\_smtp, (2, 0)) -> (pe\_schtask, (2, 0) -> (e\_http, (3, 1))

#      -> subnet\_scan -> (e\_ssh, (5, 0)) -> (e\_samba, (5, 0))

#  Score = 200 - (2+1+3+1+2+1+3+2) = 185

#

subnets: [1, 1, 5, 5, 4]

topology: [[ 1, 1, 0, 0, 0, 0],

           [ 1, 1, 1, 1, 0, 0],

           [ 0, 1, 1, 1, 0, 0],

           [ 0, 1, 1, 1, 1, 1],

           [ 0, 0, 0, 1, 1, 0],

           [ 0, 0, 0, 1, 0, 1]]

sensitive\_hosts:

  (2, 0): 100

  (5, 0): 100

os:

  - linux

  - windows

services:

  - ssh

  - ftp

  - http

  - samba

  - smtp

processes:

  - tomcat

  - daclsvc

  - schtask

exploits:

  e\_ssh:

    service: ssh

    os: linux

    prob: 0.9

    cost: 3

    access: user

  e\_ftp:

    service: ftp

    os: windows

    prob: 0.6

    cost: 1

    access: root

  e\_http:

    service: http

    os: None

    prob: 0.9

    cost: 2

    access: user

  e\_samba:

    service: samba

    os: linux

    prob: 0.3

    cost: 2

    access: root

  e\_smtp:

    service: smtp

    os: windows

    prob: 0.6

    cost: 3

    access: user

privilege\_escalation:

  pe\_tomcat:

    process: tomcat

    os: linux

    prob: 1.0

    cost: 1

    access: root

  pe\_daclsvc:

    process: daclsvc

    os: windows

    prob: 1.0

    cost: 1

    access: root

  pe\_schtask:

    process: schtask

    os: windows

    prob: 1.0

    cost: 1

    access: root

service\_scan\_cost: 1

os\_scan\_cost: 1

subnet\_scan\_cost: 1

process\_scan\_cost: 1

host\_configurations:

  (1, 0):

    os: linux

    services: [http]

    processes: []

  (2, 0):

    os: windows

    services: [smtp]

    processes: [schtask]

  (3, 0):

    os: windows

    services: [ftp]

    processes: [schtask]

  (3, 1):

    os: windows

    services: [ftp, http]

    processes: [daclsvc]

  (3, 2):

    os: windows

    services: [ftp]

    processes: []

  (3, 3):

    os: windows

    services: [ftp]

    processes: [schtask]

  (3, 4):

    os: windows

    services: [ftp]

    processes: [schtask]

  (4, 0):

    os: linux

    services: [ssh]

    processes: []

  (4, 1):

    os: linux

    services: [ssh]

    processes: []

  (4, 2):

    os: linux

    services: [ssh]

    processes: []

  (4, 3):

    os: windows

    services: [ssh, ftp]

    processes: [tomcat]

  (4, 4):

    os: windows

    services: [ssh, ftp]

    processes: [tomcat]

  (5, 0):

    os: linux

    services: [ssh, samba]

    processes: []

  (5, 1):

    os: linux

    services: [ssh, http]

    processes: [tomcat]

  (5, 2):

    os: linux

    services: [ssh]

    processes: []

  (5, 3):

    os: linux

    services: [ssh]

    processes: []

firewall:

  (0, 1): [http]

  (1, 0): []

  (1, 2): [smtp]

  (2, 1): [ssh]

  (1, 3): []

  (3, 1): [ssh]

  (2, 3): [http]

  (3, 2): [smtp]

  (3, 4): [ssh, ftp]

  (4, 3): [ftp, ssh]

  (3, 5): [ssh, ftp]

  (5, 3): [ftp, ssh]

step\_limit: 2000

1. programs

Required

To ensure that research outcomes can be validated:

* all data and/or programs which you created and/or gathered and used for the results presented in this document must be submitted;
* a softcopy of the data and/or programs must also be provided;
* if the individual file size exceeds 500 MB, please contact the Unit Coordinator on how to submit your files;
* Ensure that you use logical filenames which directly relate to a section, table or figure number or name of this document.
* If you used data and/or programs from by 3rd parties (e.g. from open access databases), you must include links to all sources used in a separate pdf document. These sources must be freely accessible to your first supervisor;
* All open access sources and databases used must also be referenced in the References section of this document;

Optional

In addition, you may include items in this appendix if that is relevant. If in doubt, please discuss with your supervisor(s).

Appendix C: Figures[[6]](#footnote-6)

*(required)*

Include in this appendix all figures used for the paper section of this document and any additional figures that may be relevant. If in doubt, ask your supervisor(s).

* Figures included in this section should have separate figure numbers from the figures used in the paper section (continue numbering your figures throughout this document) and be included in the List of Figures at the start of this document;
* Figures included in this section would normally be at a larger scale than the figures in the paper section;
* Figures should be in colour, however the information must be unambiguous in black and white (use e.g. different line types to differentiate information);
* The title of the figure should not be included in the image, but be typed below the figure;
* The title of the figure should be “Fig.” followed by a number and the text labelling the figure;
* All data and programs which were used to create figures of this document must be submitted in Learn line, refer to Appendix B for further instructions.

Appendix D: Tables[[7]](#footnote-7)

*(required if you use tables in this document)*

Include in this appendix all tables used for the paper section of this document and any additional tables that may be relevant. If in doubt, ask your supervisor(s).

* Tables included in this section should have separate table numbers from the figures used in the paper section (continue numbering your tables throughout this document) and be included in the List of Tables at the start of this document.
* Tables included in this section would normally be at a larger scale than the tables in the paper section;
* Tables should not use colour;
* The title of the table should not be included in the image, but be typed above the table;
* The title of the table should be “Table” followed by a number and the text labelling the table;
* All data and programs which were used to create tables of this document must be submitted in Learnline, refer to Appendix B for further instructions

References[[8]](#footnote-8)

[] C. Heinl, Artificial (intelligent) agents and active cyber defence: policy implications. 6th Int. Confe. on Cyber Conflict. NATO CCD COE Publications, 2016, Tallinn.

[] M. Backes, J. Hoffmann, R. Kunnemann, P. Speicher and M. Steinmetz, Simulated penetration testing and mitigation analysis. http://arxiv.org/abs/1705.05088, 2017.

[] S. Jimenez, T. De-la-rosa, S. Fernandez, F. Fernandez and D. Borrajo, A review of machine learning for automated planning. The Knowledge Engineering Review, Vol. 00:0, pp. 124. 2009. [] Y. Andrew and M. Jordan, PEGASUS: A policy search method for large MDPs and POMDPs. 16th Conf. on Uncertainty in Artificial Intel., 2013.

[] T. Schaul, J. Quan, I. Antonoglou and D. Silver, Prioritized experience replay, Google DeepMind. ICLR 2016.

[] K. Veeramachaneni, I. Arnaldo, A. Cuesta-Infante, V. Korrapati, C. Bassias and K. Li, AI2: Training a big data machine to defend. CSAIL, MIT Cambridge, 2016.

[] K. Durkota, V. Lisy, B. Bosansk and C. Kiekintveld, Optimal network security hardening using attack graph games. 24th Int. Joint Conf. on Artificial Intelligence (IJCAI-2015), 2015.

[] N. Meuleau, K. Kim, L. Kaelbling and A. Cassandra, Solving POMDPs by searching the space of finite policies. 15th Conf. on Uncertainty in Artificial Intel., 2013.

[] NIST, Computer Security Resource Center - National Vulnerability Database, https://nvd.nist.gov, 2017.

[]. Creasey, J.; Glover, I. A guide for Running an Effective Penetration Testing Program; CREST Publication: Slough, UK, 2017. Available online: http://www.crest-approved.org (accessed on 18 December 2019).

[]. Almubairik, N.; Wills, G. Automated penetration testing based on a threat model. In Proceedings of the 11th International Conference for Internet Technologies and Secured Transactions, ICITST, Barcelona, Spain, 5–7 December 2016 .

[]. Applebaum, A.; Miller, D.; Strom, B.; Korban, C.; Wol, R. Intelligent, automated red team emulation. In Proceedings of the 32nd Annual Conference on Computer Security Applications (ACSAC’16), LosAngeles, CA, USA, 5–8 December 2016; pp. 363–373.

[]. Obes, J.; Richarte, G.; Sarraute, C. Attack planning in the real world. arXiv 2013, arXiv:1306.4044.

[]. Spaan, M. Partially Observable Markov Decision Processes, Reinforcement Learning: State of the Art; Springer: Berlin/Heidelberg, Germany, 2012.

[]. Hoffmann, J. Simulated penetration testing: From Dijkstra to aaTuring Test++. In Proceedings of the 25th International Conference on Automated Planning and Scheduling, Israel, 7–11 June 2015.

[]. Sarraute, C. Automated attack planning. Available online: https://arxiv.org/abs/1307.7808 (accessed on 20 December 2019).

[]. Qiu, X.; Jia, Q.; Wang, S.; Xia, C.; Shuang, L. Automatic generation algorithm of penetration graph in penetration testing. In Proceedings of the Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Guangdong, China, 8–10 November 2014.

[]. Heinl, C. Artiﬁcial (intelligent) agents and active cyber defence: Policy implications. In Proceedings of the 6th International Conference On Cyber Conﬂict (CyCon 2014), Tallinn, Estonia, 3–6 June 2014.

[]. Sarraute, C.; Buffet, O.; Hoffmann, J. POMDPs make better hackers: Accounting for uncertainty in penetration testing. Available online: https://arxiv.org/abs/1307.8182 (accessed on 20 December 2019).

[]. Backes, M.; Hoffmann, J.; Kunnemann, R.; Speicher, P.; Steinmetz, M. Simulated Penetration Testing and Mitigation Analysis. arXiv 2017, arXiv:1705.05088.

[]. Ghanem, M.; Chen, T. Reinforcement Learning for Intelligent Penetration Testing. In Proceedings of the WS4 the World Conference on Smart Trends in Systems, Security and Sustainability, London, UK, 30–31 October 2018.

[]. J. Creasey, and I. Glover, A guide for running an effective Penetration Testing program,721 http://www.crest-approved.org. CREST Publication , 2017.722

[]. N. Almubairik, G. Wills, Automated penetration testing based on a threat model. 11th International723 Conference for Internet Technologies and Secured Transactions, ICITST, 2016.724

[]. A. Applebaum, D. Miller, B. Strom, C. Korban, and R. Wol, Intelligent, automated red team emulation. 32nd725 Annual Conference on Computer Security Applications (ACSAC ’16), 2016, pp. 363-373.726

[]. J. Obes, G. Richarte, and C. Sarraute, Attack planning in the real world. Journal CoRR Article, 2013,727 abs/1306.4044.728

[]. M. Spaan, Partially Observable Markov Decision Processes, Reinforcement Learning: State of the Art,729 Springer Verlag, 2012.730

[]. J.Hoffmann,Simulatedpenetrationtesting: FromDijkstratoaaTuringTest++. 25thInt. Conf. onAutomated731 Planning and Scheduling, 2015, AAAI Press.732

[]. C.Sarraute,Automatedattackplanning. InstitutoTecnologicodeBuenos-aires,Ph.D.Thesis,2012,Argentina.733

[]. X. Qiu, Q. Jia, S. Wang, C.Xia and L. Shuang, Automatic generation algorithm of penetration graph in734 penetration testing, 19th Int. Conf. on P2P, Parallel, Grid, Cloud and Internet Computing, 2014.735

[]. C. Sarraute, O. Buffet and J. Hoffmann, POMDPs make better hackers: Accounting for uncertainty in736 penetration testing. 26th AAAI Conf. on Artiﬁcial Intel. (AAAI’12), pp. 1816–1824, July 2012.737

[]. C. Heinl, Artiﬁcial (intelligent) agents and active cyber defence: policy implications. 6th Int. Confe. on738 Cyber Conﬂict. NATO CCD COE Publications, 2016, Tallinn.739

[]. M. Backes, J. Hoffmann, R. Kunnemann, P. Speicher and M. Steinmetz, Simulated penetration testing and740 mitigation analysis. http://arxiv.org/abs/1705.05088, 2017.741

[]. S. Jimenez, T. De-la-rosa, S. Fernandez, F. Fernandez and D. Borrajo, A review of machine learning for742 automated planning. The Knowledge Engineering Review, Vol. 00:0, pp. 1–24. 2009.743

[]. Y. Andrew and M. Jordan, PEGASUS: A policy search method for large MDPs and POMDPs. 16th Conf. on744 Uncertainty in Artiﬁcial Intel., 2013.745

[]. T. Schaul, J. Quan, I. Antonoglou and D. Silver, Prioritized experience replay, Google DeepMind. ICLR 2016.746

[]. K.Veeramachaneni,I.Arnaldo,A.Cuesta-Infante,V.Korrapati,C.BassiasandK.Li,AI2: Trainingabigdata747 machine to defend. CSAIL, MIT Cambridge, 2016.748

[]. K. Durkota, V. Lisy, B. Bosansk and C. Kiekintveld, Optimal network security hardening using attack graph749 games. 24th Int. Joint Conf. on Artiﬁcial Intelligence (IJCAI-2015), 2015.750

[]. N. Meuleau, K. Kim, L. Kaelbling and A. Cassandra, Solving POMDPs by searching the space of ﬁnite751 policies. 15th Conf. on Uncertainty in Artiﬁcial Intel., 2013.752

[]. NIST, Computer Security Resource Center - National Vulnerability Database, https://nvd.nist.gov, 2018.753

[]. E. Walraven, amd M. Spaan. Accelerated Vector Pruning for Optimal POMDP Solvers, Proceedings of the754 Thirty-First AAAI Conference on Artiﬁcial Intelligence, 2017.755

[]. M. Ghanem, and T. Chen. Reinforcement Learning for Intelligent Penetration Testing. World Conference on756 Smart Trends in Systems, Security and Sustainability. 2018.

[] Symantec Corporation, “Internet Security Threat Report,” Volume 22, Apr. 2017.

[] Australian Cyber Security Centre, “ACSC Threat Report,” Australian Government, 2017.

[] A. IVan, “Why Attacking Systems Is a Good Idea,” ​IEEE Secur. Priv. ​ , 2004.

[] R. S. Sutton and A. G. Barto, “Reinforcement learning: An introduction,” 2011.

[] D. of Industry, “$50 million investment into cyber security research and industry solutions,” Ministers for the Department of Industry, Innovation and Science ​ , 22-Sep-2017. [Online]. Available: http://minister.industry.gov.au/ministers/craiglaundy/media-releases/50-million-investment-cyber-se curity-research-and-industry​. [Accessed: 13-Mar-2018].

[] B. Arkin, S. Stender, and G. McGraw, “Software penetration testing,” ​IEEE Secur. Priv. ​ , vol. 3, no. 1, pp. 84–87, Jan.-Feb 2005.

[]Hansen, E. A., and Feng, Z. 2000. Dynamic programming for POMDPs using a factored state representation. In Proceedings of the 5th international conference on Artificial Intelligence Planning & Scheduling.

[]Cassandra, A.; Littman, M.; and Zhang, N. 1997. Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In Proceedings of the 13th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-97), 54–61.

1. List here all figures you have used in the paper and other sections of this document, number all figures, remove this footnote [↑](#footnote-ref-1)
2. List here all tables you have used in the paper and other sections of this document, number all tables, remove this footnote [↑](#footnote-ref-2)
3. List here all abbreviations used in the paper and the other sections of this document, remove this footnote [↑](#footnote-ref-3)
4. List here all variables and their units used in the paper and other sections of this document, remove this footnote [↑](#footnote-ref-4)
5. Amend headings and text as needed, remove this footnote [↑](#footnote-ref-5)
6. Amend headings and text as needed, remove this footnote [↑](#footnote-ref-6)
7. Amend headings and text as needed, remove this footnote [↑](#footnote-ref-7)
8. It is highly recommended that you use EndNote to make the list of references for this document, remove this footnote [↑](#footnote-ref-8)